

DESIGN AND ANALYSIS OF CLINICAL TRIALS

Second Edition

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie,
Nicholas I. Fisher, Iain M. Johnstone, J. B. Kadane, Geert Molenberghs, Louise M. Ryan,
David W. Scott, Adrian F. M. Smith, Jozef L. Teugels;*
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, David G. Kendall*

A complete list of the titles in this series appears at the end of this volume.

DESIGN AND ANALYSIS OF CLINICAL TRIALS

Concepts and Methodologies

Second Edition

SHEIN-CHUNG CHOW

Millennium Pharmaceuticals, Inc.
Cambridge, MA

JEN-PEI LIU

National Cheng-kung University
Tainan, Taiwan

National Health Research Institutes
Taipei, Taiwan

 **WILEY-INTERSCIENCE**

A John Wiley & Sons, Inc., Publication

Copyright © 2004 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-646-8600, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data is available.

ISBN 0-471-24985-8

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

Preface	ix
Preface to the First Edition	xi
1. Introduction	1
1.1 What are Clinical Trials?, 1	
1.2 History of Clinical Trials, 3	
1.3 Regulatory Process and Requirements, 7	
1.4 Investigational New Drug Application, 15	
1.5 New Drug Application, 22	
1.6 Clinical Development and Practice, 31	
1.7 Aims and Structure of the Book, 35	
2. Basic Statistical Concepts	43
2.1 Introduction, 43	
2.2 Uncertainty and Probability, 44	
2.3 Bias and Variability, 47	
2.4 Confounding and Interaction, 55	
2.5 Descriptive and Inferential Statistics, 65	
2.6 Hypothesis Testing and <i>p</i> -Values, 71	
2.7 Clinical Significance and Clinical Equivalence, 77	
2.8 Reproducibility and Generalizability, 82	
3. Basic Design Considerations	88
3.1 Introduction, 88	
3.2 Goals of Clinical Trials, 89	

3.3 Target Population and Patient Selection, 93	
3.4 Selection of Controls, 100	
3.5 Statistical Considerations, 109	
3.6 Other Issues, 116	
3.7 Discussion, 118	
4. Randomization and Blinding	120
4.1 Introduction, 120	
4.2 Randomization Models, 122	
4.3 Randomization Methods, 127	
4.4 Implementation of Randomization, 149	
4.5 Generalization of Controlled Randomized Trials, 154	
4.6 Blinding, 158	
4.7 Discussion, 165	
5. Designs for Clinical Trials	167
5.1 Introduction, 167	
5.2 Parallel Group Designs, 169	
5.3 Cluster Randomized Designs, 174	
5.4 Crossover Designs, 179	
5.5 Titration Designs, 188	
5.6 Enrichment Designs, 194	
5.7 Group Sequential Designs, 200	
5.8 Placebo-Challenging Design, 202	
5.9 Blinded Reader Designs, 208	
5.10 Discussion, 212	
6. Designs for Cancer Clinical Trials	215
6.1 Introduction, 215	
6.2 General Considerations for Phase I Cancer Clinical Trials, 217	
6.3 Single-Stage Up-and-Down Phase I Designs, 218	
6.4 Two-Stage Up-and-Down Phase I Designs, 220	
6.5 Continual Reassessment Method Phase I Designs, 223	
6.6 Optimal/Flexible Multiple-Stage Designs, 226	
6.7 Randomized Phase II Designs, 232	
6.8 Discussion, 236	
7. Classification of Clinical Trials	239
7.1 Introduction, 239	
7.2 Multicenter Trial, 240	
7.3 Superiority Trials, 247	
7.4 Active Control and Equivalence/Noninferiority Trials, 250	
7.5 Dose-Response Trials, 265	
7.6 Combination Trials, 270	
7.7 Bridging Studies, 283	
7.8 Vaccine Clinical Trials, 289	
7.9 Discussion, 296	

8. Analysis of Continuous Data	300
8.1 Introduction, 300	
8.2 Estimation, 301	
8.3 Test Statistics, 305	
8.4 Analysis of Variance, 311	
8.5 Analysis of Covariance, 316	
8.6 Nonparametrics, 320	
8.7 Repeated Measures, 326	
8.8 Discussion, 337	
9. Analysis of Categorical Data	339
9.1 Introduction, 339	
9.2 Statistical Inference for One Sample, 344	
9.3 Inference of Independent Samples, 356	
9.4 Ordered Categorical Data, 362	
9.5 Combining Categorical Data, 366	
9.6 Model-Based Methods, 372	
9.7 Repeated Categorical Data, 379	
9.8 Discussion, 384	
10. Censored Data and Interim Analysis	386
10.1 Introduction, 386	
10.2 Estimation of the Survival Function, 388	
10.3 Comparison between Survival Functions, 394	
10.4 Cox's Proportional Hazard Model, 402	
10.5 Calendar Time and Information Time, 417	
10.6 Group Sequential Methods, 422	
10.7 Discussion, 435	
11. Sample Size Determination	438
11.1 Introduction, 438	
11.2 Basic Concept, 439	
11.3 Two Samples, 443	
11.4 Multiple Samples, 452	
11.5 Censored Data, 464	
11.6 Dose-Response Studies, 468	
11.7 Crossover Designs, 474	
11.8 Equivalence and Noninferiority Trials, 481	
11.9 Multiple-Stage Design in Cancer Trials, 492	
11.10 Comparing Variabilities, 493	
11.11 Discussion, 508	
12. Issues in Efficacy Evaluation	510
12.1 Introduction, 510	
12.2 Baseline Comparison, 512	

12.3 Intention-to-Treat Principle and Efficacy Analysis,	517
12.4 Adjustment for Covariates,	523
12.5 Multicenter Trials,	529
12.6 Multiplicity,	537
12.7 Data Monitoring,	546
12.8 Use of Genetic Information for Evaluation of Efficacy,	552
12.9 Sample Size Re-estimation,	558
12.10 Discussion,	560
13. Safety Assessment	562
13.1 Introduction,	562
13.2 Extent of Exposure,	564
13.3 Coding of Adverse Events,	569
13.4 Analysis of Adverse Events,	584
13.5 Analysis of Laboratory Data,	591
13.6 Discussion,	600
14. Preparation and Implementation of a Clinical Protocol	602
14.1 Introduction,	602
14.2 Structure and Components of a Protocol,	603
14.3 Points to Be Considered and Common Pitfalls during Development and Preparation of a Protocol,	609
14.4 Common Departures for Implementation of a Protocol,	612
14.5 Monitoring, Audit, and Inspection,	617
14.6 Quality Assessment of a Clinical Trial,	620
14.7 Discussion,	626
15. Clinical Data Management	628
15.1 Introduction,	628
15.2 Regulatory Requirements,	630
15.3 Development of Case Report Forms,	633
15.4 Database Development,	636
15.5 Data Entry, Query, and Correction,	638
15.6 Data Validation and Quality,	641
15.7 Database Lock, Archive, and Transfer,	642
15.8 Discussion,	645
Bibliography	649
Appendices	683
Index	713

PREFACE

In recent years, there has been an explosive growth of literature in clinical trials. As indicated in the first edition, the purpose of this book is to provide a comprehensive and unified presentation of the principles and methodologies in designs and analyses utilized for various clinical trials and to give a well-balanced summary of current regulatory requirements and recently developed statistical methods in this area. Since the first edition was published in 1998, it has been well received by clinical scientists/researchers and is now widely used as a reference source and a graduate textbook in clinical research and development. It is our continuing goal to provide a complete, comprehensive, and updated reference and textbook in the area of clinical research.

The second edition can be distinguished from the first in three ways. First, we have revised and/or updated sections to reflect good clinical practice in regulatory review/approval process and recent developments in design and analysis in clinical research. For example, the second edition provides an update of the status of clinical trials and regulations, especially ICH (International Conference on Harmonization) guidelines for clinical trials since 1998. Second, the second edition is expanded to 15 chapters. Additional new topics and three new chapters are added to provide a total account of the most recent development in clinical trials. To name just a few, the second edition includes new topics such as clinical significance and reproducibility and generalizability (Chapter 2); goals of clinical trials and target population (Chapter 4); clustered randomized design, group sequential design, placebo-challenging design, and blinded reader design (Chapter 5); superiority trials, active control and equivalence/noninferiority trials, dose-response trials, bridging studies, and vaccine clinical trials (Chapter 7); sample size determination on equivalence and noninferiority trials and comparing variabilities (Chapter 11); and use of genomic information for evaluation of efficacy (Chapter 12). The three new chapters include “Designs for Cancer Clinical Trials” (Chapter 6), “Preparation and Implementation of a Clinical Protocol” (Chapter 14), and “Clinical Data Management” (Chapter 15).

Finally, the second edition includes more than 280 new references from clinical-related literature. We believe that this revised and expanded second edition will benefit clinical scientists/researchers from the medical-pharmaceutical industry, regulatory agencies, and academia by serving as an extremely useful reference source in clinical research.

From John Wiley and Sons, I would like to thank Steve Quigley for providing us the opportunity to work on this edition, and Susanne Steitz for her outstanding efforts in preparing this edition. The first author would like to thank support from colleagues from StarPlus, Inc. and Millennium Pharmaceuticals, Inc. during the preparation of this edition. The second author wishes to express his gratitude to his wife, Dr. Wei-Chu Chie, and their daughter Angela for their support, patience, and understanding during the preparation of this edition.

Finally, the views expressed are those of the authors and not necessarily those of Millennium Pharmaceuticals, Inc., and National Cheng-Kung University and National Health Research Institutes, Taiwan. We are solely responsible for the contents and errors of this edition. Any comments and suggestions will be very much appreciated.

SHEIN-CHUNG CHOW
JEN-PEI LIU

*Cambridge, Massachusetts
Tainan, Taiwan
September, 2003*

PREFACE TO THE FIRST EDITION

Clinical trials are scientific investigations that examine and evaluate safety and efficacy of drug therapies in human subjects. Biostatistics has been recognized and extensively employed as an indispensable tool for planning, conduct, and interpretation of clinical trials. In clinical research and development, the bio-statistician plays an important role that contributes toward the success of the trial. An open and effective communication among clinician, biostatistician, and other related clinical scientists will result in a successful clinical trial. The mutual communication, however, is a two-way street: not only (1) the biostatistician must effectively deliver statistical concepts and methodologies to his/her colleagues but also (2) the clinician must communicate thoroughly clinical and scientific principles embedded in clinical research to the biostatistician. The biostatistician can then formulate these clinical and scientific principles into valid statistical hypotheses, models, and methodologies for data analyses. The integrity, quality, and success of a clinical trial depend on the interaction, mutual respect, and understanding among the clinician, the biostatistician, and other clinical scientists.

There are many books on clinical trials already on the market. These books, however, emphasize either statistical or clinical aspects. None of these books provides a balanced view of statistical concepts and clinical issues. Therefore the purpose of this book is not only to fill the gap between clinical and statistical disciplines but also to provide a comprehensive and unified presentation of clinical and scientific issues, statistical concepts, and methodologies. Moreover this book focuses on the interactions among clinicians, biostatisticians, and other clinical scientists that often occur during the various phases of clinical research and development. This book is intended to give a well-balanced overview of current and emerging clinical issues and newly developed statistical methodologies. Although this book is written from a viewpoint of pharmaceutical research and development, the principles and concepts presented in this book can be applied to nonbiopharmaceutical settings.

It is our goal to provide a concise and comprehensive reference book for physicians, clinical researchers, pharmaceutical scientists, clinical or medical research associates, clinical programmers or data coordinators, and biostatisticians in the areas of clinical research and development, regulatory agencies, and academe. Hence this book is written for readers with minimal mathematical and statistical backgrounds. Although it is not required, an introductory statistics course that covers the concepts of probability, sampling distribution, estimation, and hypothesis testing would be helpful. This book can also serve as a textbook for graduate courses in the areas of clinical and pharmaceutical research and development. Readers are encouraged to pay attention to clinical issues and their statistical interpretations as illustrated through real examples from various phases of clinical research and development.

The issues covered in this book may occur during the various phases of clinical trials in pharmaceutical research and development, and their corresponding statistical interpretations, concepts, designs, and analyses. All the important clinical issues are addressed in terms of the concepts and methodologies of the design and analysis of clinical trials. For this reason this book is composed of clinical concepts and methodologies. Each chapter with different topics is self-contained.

Chapter 1 provides an overview of clinical development for pharmaceutical entities, the process of drug research and development in pharmaceutical industry, and regulatory processes and requirements. The aim and structure of the book is also discussed in this chapter. The concepts of design and analysis of clinical trials are covered from Chapters 2 through 6. Basic statistical concepts such as uncertainty, bias, variability, confounding, interaction, and statistical versus clinical significance are introduced in Chapter 2. Fundamental considerations for the selection of a suitable design in achieving certain objectives of a particular trial under various circumstances are provided in Chapter 3. Chapter 4 illustrates the concepts and different methods of randomization and blinding that are indispensable to the success and integrity of a clinical trial. Chapter 5 introduces different types of statistical designs for clinical trials such as parallel, crossover, titration, and enrichment designs and discusses their relative advantages and drawbacks. Various types of clinical trials, which include multicenter, active control, combination, and equivalence trials, are the subject of Chapter 6.

Methodologies and the issues for clinical data analysis are addressed in Chapters 7 through 12. Since clinical endpoints can generally be classified into three types, continuous, categorical, and censored data, various statistical methods for analyses of these three types of clinical data and their advantages and limitations are provided in Chapters 7, 8, and 9, respectively. In addition, group sequential procedures for interim analysis are given in Chapter 9. Different procedures for sample size determination are provided in Chapter 10 for data under different designs. Statistical issues in analyzing efficacy data are discussed in Chapter 11. These issues include baseline comparisons, intention-to-treat analyses versus evaluable or per-protocol analyses, adjustment of covariates, multiplicity issues, and data monitoring. Chapter 12 focuses on the issues of analysis of safety data, including the extent of exposure, coding, and analysis of adverse events, and analysis of laboratory data.

For each chapter, whenever possible, real examples from clinical trials are included to demonstrate the clinical and statistical concepts, interpretations, and their relationships and interactions. Comparisons of the relative merits and disadvantages of statistical methodology for addressing different clinical issues in various therapeutic areas are discussed in appropriate chapters. In addition, if applicable, topics for future development are provided.

All computations in this book were performed using SAS. Other statistical packages such as SPSS, BMDP, or MINTAB may also be applied.

At John Wiley, we would like to thank Acquisition Editor Steve Quigley for providing us with the opportunity to work on this book and for his outstanding effort in preparing this book for publication. We are greatly indebted to the Bristol-Myers Squibb Company and Covance, Inc. for their support, in particular, to S. A. Henry, L. Meinert, and H. Koffer. We are grateful for A. P. Pong, C. C. Hsieh, and G. Y. Han for their assistance in preparing the many charts, figures, graphs, and tables in this book. We are grateful to Y. C. Chi, F. Ki, and C. S. Lin for many helpful discussions and for reviewing the manuscript. We also wish to thank A. P. Pong, M. L. Lee, and E. Nordbrock for their constant support and encouragement. The first author also wishes to express his appreciation to his wife, Yueh-Ji, and their daughters, Emily and Lilly, for their patience and understanding during the preparation of this book.

Finally, we are fully responsible for any errors remaining in the book. The views expressed are those of the authors and are not necessarily those of Covance, Inc. and the National Cheng-Kung University.

SHEIN-CHUNG CHOW
JEN-PEI LIU

Princeton, New Jersey

Tainan, Taiwan

October 1997

INDEX

- Absorbing event, 560
Absorption, distribution, metabolism, and excretion (ADME), 15, 16
Accelerated approval, 6, 27, 554
Accuracy, 47, 93, 98, 126
Additivity, 371
Adherence, 117
Adverse drug experience, 570
Adverse drug reaction (ADR), 569, 573
 serious (SADR), 573
 unexpected (UADR), 573
Adverse event (AE), 19, 91, 110, 564, 569–572, 574, 608
 absorbing, 565
 common, 584, 586
 primary, 584
 rare, 586
 secondary, 584
 serious (SAE), 19, 59, 215, 362, 511, 564, 574, 584, 608
 significant, 564, 569, 585
 treatment-emergent (TEAE), 572
Adverse Event Expedited Reporting System (AdEERS), 584
Adverse experience, 21, 569
 serious, 21
 unexpected, 21
Adverse reaction, 15, 117, 563
Advisory committee, 30, 51
AIDS, 31, 43, 100, 103, 256
Alternative
 fixed local, 465
 Lehmann's, 405
 one-sided, 428, 431
 two-sided, 535
Alzheimer's disease assessment scale (ADAS), 195–196, 607
Analysis
 as-treated, 259
 change from baseline, 353
 cross-sectional, 333
 efficacy, 511, 520
 evaluable, 518, 520
 intention-to-treat (ITT), 59, 110, 257, 419, 512, 518–520, 624
 interim, 19, 89, 109, 112–114, 201, 298, 386–387, 424, 511–512, 609
 administrative, 549–550
 formal, 549
 unplanned, 547
 unreported, 547
meta-, 115–116, 166, 247, 285
ordinary regression, 408
per-protocol, 520
power, 439, 443
precision, 439, 441
preferred, 520
prestudy power, 439

- regression, 527
- shift, 595
- stratified, 141, 431
 - post-treatment, 535
- subgroup, 145, 511, 538, 545–546
- survival, 389, 464, 588
- Analysis set, 609
 - full, 518
- Analysis of covariance (ANCOVA), 118, 316, 319, 527, 538, 546
- Analysis of variance (ANOVA), 311, 337, 527
 - one-way, 337, 452
 - two-way, 552
- Analysis of variance table, 312, 316, 541
- Angiotensin converting enzyme (ACE) inhibitor, 7, 191, 551
- Approach
 - alpha spending, 423
 - backwards, 633
 - Bayesian, 288
 - confidence interval, 306, 442
 - maximum error, 442
 - population-average, 337
 - quasi-likelihood, 334
- Approximation, 47
 - large-sample, 365
 - non-central chi-square, 463
 - normal, 324
- Arrhythmia, 96
 - ventricular, 515
- Aspirin, 60, 62
- Assumption
 - consistency, 263
 - fundamental bioequivalence, 115–116
 - normality, 371, 528, 543
 - proportional hazard, 408
- Atrial fibrillation, 92
- Average, 55, 155
 - adjusted treatment, 526
 - overall, 525
 - population, 70, 534
 - sample, 70
 - simple, 534
 - unadjusted treatment, 526
- AUA-7 symptom score, 340, 537
- Audit, 617
- Audit trail, 150, 630
- Baseline, 17, 65, 125, 307, 318, 513
 - multiple, 514
- Baseline characteristic, 19, 66, 509
- Baseline, comparability, 511, 512
- Baseline hazard, 405
- Baseline value, 512–514
- Beta-block, 191
- Beta-blocker Heart Attack Trial (BHAT), 164, 390, 418, 421, 430
- Bias, 23, 43, 47, 48, 51, 89, 93, 97, 100, 112, 120, 129, 140, 437
 - accidental, 48, 128–129, 135–136, 141
 - investigator's, 128
 - judgmental, 48
 - observer's, 102
 - operational, 48
 - selection, 17, 48, 51, 127–128, 135, 140
 - statistical, 48
 - subjective, 127
- Bioavailability, 3
- Bioequivalence, 3, 80–81, 436
 - individual, 158, 497
 - population, 158, 497
- Bioinformatics, 511–512
- Biologic Act, 12
- Biological product, 3, 8, 11
- Biostatistician, 35, 119, 149, 156, 610
- Biostatistics, 31, 35
- Blind
 - double, 23, 105, 159, 160, 173, 192
 - single, 159–160, 173
 - triple, 23, 113, 159–160
- Blinding, 17, 89, 118, 120, 158, 165
 - double, 121, 239
 - single, 121
 - triple, 121
- Blindness, 113, 158
- Blocking, 54, 113, 121
- Blood chemistry, 150, 608
- Body System Classification, 575
- Bone mineral density (BMD), 54, 90
- Bonferroni adjustment, 539, 542
- Bonferroni correction, 542
- Bonferroni technique, 539, 542
- Boundary
 - asymmetric
 - lower, 434
 - upper, 433
 - group sequential, 423–424, 426
 - O'Brien—Fleming, 425–426, 434
 - O'Brien—Fleming group sequential, 426
 - one-sided, 431
 - Pocock's group sequential, 425
- British Medical Research Council, 120
- B-value, 388, 432–433
- Cancer Therapy Evaluation Program (CTEP), 574
- Cardiac arrhythmia, 97
- Cardiac Arrhythmia Suppression Trial (CAST), 45, 103, 189, 196, 199, 211, 418, 431, 432
- Case report form, 33, 42, 65, 119, 212, 618, 628, 631
 - design, 638
 - flow, 633–634
 - principles, 633
 - tracking, 633–634

- tracking system, 633
- transmission form, 634–635
- Case report tabulation, 511
- Center, 246, 602
 - Center for Biologics Evaluation and Research (CBER), 8, 11, 13, 35, 552
 - Center for Devices and Radiological Health (CDRH), 8, 11, 13
 - Center for Drug Evaluation and Research (CDER), 4, 8, 11, 13, 35
 - Center for Food Safety and Applied Nutrition (CFSAN), 8
 - Center for Veterinary Medicine (CVM), 8
- Central clinical trial file, 634
- Central limit theorem, 47, 303, 333
- Central tendency, 46
- Change from the baseline, 307–308, 319, 423, 512, 538
 - absolute, 512
 - percent, 512
- Chebyshev inequality, 442
- Cholesterol, 46
 - Cholesterol and Recurrent Events Trial, 402
- Chronic lymphocytic leukemia (CLL), 27
- Chronic myeloid leukemia (CML), 27
- Chronic obstructive pulmonary disease (COPD), 91
- Classification
 - one-way, 311, 319
 - two-way, 315
- Clinical Data Update System (CDUS), 584
- Clinical Development Plan, 31–32
- Clinical Development Program, 31
- Clinical investigator, 14
 - phase I, 14
- Clinical monitor, 159
- Clinical research, 36
- Clinical research associate, 150, 160
- Clinical significant change, 593
- Clinical trials
 - audit, 617–619
 - definition, 1–2
 - fraud, 614–616
 - goal, 89
 - history, 3–5
 - inspection, 617
 - misconduct, 614–616
 - monitor, 617–618
- Clinician, 36, 156
- Cluster, 174
- Cluster size, 175
- Code of Federal Regulations (CFR), 1, 11, 15, 22
- Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART), 563, 575–577
- Combination
 - optimal dose, 282, 283
- Combination therapy, 270, 277, 537
- Combined product, 13, 270
- Combined therapy, 13
- Committee for Proprietary Material Products, 32, 100
 - (CPMP) Working Party, 511, 519
- Comparative clinical significance, 78
- Comparison
 - baseline, 512
 - between-group, 356
 - indirect confidence interval (ICIC), 264
 - marginal, 354
 - multiple, 19, 313, 509, 539–540
 - pairwise, 417
 - unbiased, 517
 - virtual, 264
 - within-patient, 179
 - within-treatment, 307
- Complete clinical data package (CCDP), 287
- Compliance, 59, 118, 171
 - patient, 89, 117–118
- Compliance Program Guidance Manual, 619, 632
- Composite clinical index, 340
- Composite efficacy measure, 294
- Composite index, 543
- Confidence interval, 66, 68, 70, 75, 81, 257, 303, 306–307, 309, 566
 - Bonferroni simultaneous, 313–314
 - exact, 345
 - large-sample, 393, 399, 407
 - multi-t, 313
 - repeated, 388, 426
 - Satterwaite's, 310
 - simultaneous, 312, 541
 - symmetric, 70
 - two-sided, 247
- Confidence limit
 - one-sided, 81, 259
- Confounding, 43, 55, 57
- Congestive heart failure (CHF), 201
- Consistency, 54, 287
- Consolidated Standards for Reporting of Trials (CONSORT) group, 623
 - checklist, 625
 - statement, 624
- Contingency table
 - 2 x 2, 360, 400, 588
- Continuity correction, 355, 362, 588
- Contract Research Organization (CRO), 550, 619, 642
- Contrast
 - linear, 408
 - orthogonal, 514
 - within-subject, 182
- Control, 1, 6, 17
 - active, 100, 169, 239–240, 250–251, 255–256, 440
 - active concurrent, 22, 101, 106–107, 250
 - active treatment concurrent, 101
 - concurrent, 101, 115, 239
 - dose-comparison concurrent, 22
 - dose-response concurrent, 101, 105, 250
 - external, 101, 250

- historical, 22, 101, 109, 250, 263
- matched, 161
- no treatment, 22, 101
- no treatment current, 101, 107–108, 250
- placebo, 101, 118, 169
- placebo concurrent, 22, 100–107, 250, 263
- placebo putative, 263
- positive, 161, 251
- Control group, 165
- Cooperative North Scandinavian Enalapril Survival Study II (CONSENSUS II), 551
- Correlation
 - intrasubject, 333
 - pairwise, 544
 - within-subject, 334, 463
- Correlation coefficient, 563
 - intrablock, 142
 - intraclass (ICC), 175, 177
- Correlation matrix, within-group, 543
 - working, 334
- Coronary Drug Project Research, 153
- Cost-benefit analysis, 2
- Cost-effectiveness, 2, 31
- Cost-minimization, 2
- Covariance, 334
- Covariance matrix, 335, 373, 596
 - estimated large-sample, 407
 - sample, 382, 514
 - within-group sample, 515
 - pooled, 516
- Covariance structure, 335, 436, 505
- Covariate, 19, 60, 126, 320, 329, 388, 403, 463, 523, 609
 - adjustment, 512, 524
 - categorical, 528
 - stratified, 147
 - subject-specific, 528
 - time-dependent, 408, 413, 566, 609
 - time-independent, 403, 413, 609
- Criterion, 94
 - discontinuation, 606
 - eligibility, 94–97, 99
 - entry, 512
 - exclusion, 17, 94–97, 114, 122, 130, 611
 - inclusion, 17, 94–97, 114, 122, 130, 611
- Critical value, 74, 423
- Data
 - baseline, 513
 - binary, 359, 422
 - categorical, 300, 339, 372
 - combined, 366
 - ordered, 300, 340, 362
 - ordinal, 377
 - repeated, 341, 379, 395
 - censored, 386–390, 435, 464
 - correlated, 435–436
 - left-, 436
 - right-, 436–437
 - continuous, 301
 - cross-section, 193
 - demographic, 512
 - efficacy, 118, 548
 - genotypic, 553
 - historical, 109
 - laboratory, 569
 - longitudinal, 189, 326, 333, 337
 - measurement, 300
 - missing, 19, 511
 - numerical, 320
 - postrandomization, 518
 - ranked, 300
 - repeated, 383
 - safety, 548
- Data correction, 638, 641
- Data edit check specifications, 637–638
- Data entry, 628, 638
 - double, 638–639
- Data quality, 639, 641, 645
- Data query, 628
- Data validation, 636, 641
- Data verification, 538–641, 642, 644
- Data and safety monitoring committee (DSMC), 200, 201, 604
- Data capture
 - electronic, 645
- Data coordination and statistical analysis center, 550
- Data coordinator, 149–150, 159–160, 520
- Data management, 141, 151
 - clinical (CDM), 628
- Data management file, 645
 - master, 645
- Data management process, 647
- Data management system, 550
- Data monitoring, 89, 109, 112, 387, 511, 546–547
- Data monitoring board, 113
 - external, 113, 114
 - internal, 113
- Data Monitoring Committee (DMC), 421, 550–551
 - independent, 200, 550
- Data monitoring process, 422
- Database, 65
 - archive, 642
 - design, 636
 - development, 636
 - finalization, 642
 - lock, 642
 - transfer, 642–645
- Database development, 628
- Database management system, 550
- Dataset
 - evaluable, 511
 - intention-to-treat, 511, 523
- Death per patient year (DPPY), 565

- De-escalation, 218
Degree of freedom, 303–309, 312–316, 325, 331, 398, 452, 479
Department of Health and Human Service (DHHS), 34
Design, 88, 89
 add-on, 213
 Balaam, 181, 183, 480
 balanced incomplete block, 184–185
 blinded reader, 169, 208
 complete crossover, 177
 completely binomial, 131
 complete randomized, 169
 cluster randomized, 169, 174, 176
 crossover, 54, 58–59, 127, 169, 179, 184, 186, 474
 four-sequence, two-period (4 x 2), 203
 higher-order, 59, 181, 479
 replicated, 182, 494, 497, 504
 two-sequence, two-period, 435–436, 476
 2 x 3, 181
 crossover dose-response, 189, 266, 268
 dose escalation, 216, 268
 standard, 218
 dual, 182
 enrichment, 169, 194, 200, 512, 555
 experimental, 17, 19
 extended-period, 182
 factorial, 6, 282, 296
 2 x 2, 60, 274, 276
 2 x 2 x 2, 276
 fractional, 296
 full, 2, 274, 297
 incomplete, 296
 disconnected, 297
 multilevel, 275
 partial, 277
 flexible dose-escalation, 194
 flexible two-stage, 227
 forced dose-escalation, 169, 189, 192
 four-period, 480
 Gehan's phase II, 232
 group comparison, 169
 group randomized, 174
 group sequential, 169, 200
 matched pairs parallel, 169, 170
 minimax, 226, 232
 multiple-stage, 492
 Fleming's, 232
 flexible, 216, 226–227
 optimal, 216, 226–228
 optimal, 167, 181, 183
 optimal three-stage, 228
 optimal two-stage, 226
 orthogonal Latin squares, 184
 parallel, 54, 112
 three-group, 169, 174
 two-group, 167, 187
 parallel dose-response, 189, 266–267
 parallel-group, 169, 171–172, 186
 placebo challenging, 169, 202–203, 266, 493
 randomized block, 179
 randomized complete block, 170
 randomized incomplete block, 182
 randomized phase II, 232
 randomized parallel dose-response, 266
 randomized withdrawal, 213
 replacement, 213
 replicated, 182
 selection, 612
 single-stage, 128
 up-and-down, 218
 titration, 169, 188, 193
 accelerated, 222
 forced, 169, 266, 268
 optimal, 269
 optional, 189, 266, 269
 standard, 188
 two-period, two-sequence (2 x 2) crossover, 181, 187
 double standard, 182
 two-sequence dual, 182, 480
 two-stage, 220, 226
 flexible, 227
 Simon's optimal, 226
 up-and-down, 220
 two-stage Simon, 226
 up-and-down, 216, 218, 220
 variance-balance, 183
 Williams, 183, 266, 277
Design matrix, 463
Diabetes Prevention Program, 173
Diabetes Prevention Program Research Group, 173–174
Diagonal matrix, 590
Dictionary
 COSTART, 563, 575–577
 CTC, 574–575
 HARTS, 575
 ICD, 577
 ICD-9, 577
 ICD-9-CM, 577
 IMT, 563, 583
 J-ART, 575, 577
 MedDRA, 563, 577
 WHOART, 575–577
Difference
 clinical, 44
 clinically
 important, 80
 meaningful, 110, 112, 438, 440
 scientific meaningful, 438
 significant, 78, 80–81
 period, 475
 significant, 124
 statistical, 44
 statistical significant, 78–80, 110, 440
 treatment, 81

- Disease, 2
 Alzheimer, 110
 coronary heart (CHD), 434
- Disease incidence, 292
 very low, 293
- Dispersion, 46
- Distribution
 binomial, 351
 product, 372
 Cauchy, 513
 chi-square, 378, 380–382. See also Distribution, χ^2
 central, 380–382, 398, 400
 noncentral, 490
 exponential, 235
 extreme-value, 374
 F, 312, 316, 318, 331, 452, 494
 noncentral, 453
 hypergeometric, 361, 363, 397
 product, 370
 independent and identical, 122
 normal, 46, 302, 333, 532
 multivariate, 373, 463, 514
 permutation, 146, 487
 joint, 431
 Poisson, 384
 population, 46, 65
 prior, 224
 probability, 122, 124–125, 333
 sampling, 47, 65, 73, 125, 431
 standard normal, 321, 345, 351, 399, 423, 442, 471
 Student t-, 303, 305, 309
 t-, 313, 475, 479
 central, 477–478
 noncentral, 446
 χ^2 , 325
 central, 373
- DNA bank, 558
- Dose
 maximum effective, 173
 maximum tolerable (MTD), 41, 104, 188, 216–217, 224, 236, 266–267, 271, 540
 maximum useful (MUD), 266
 minimum effective (MED), 105, 188, 266–267, 468, 470, 540
- Dose-response, 30, 265
- Dose-response curve, 435
 individual average (IDRC), 265
 population average (PDRC), 265
- Dosing range, 14
- Double-dummy, 162
- Dropout, 89, 117, 511, 560
- Drug, 3, 8, 11
 combination, 77, 271, 273, 278
 fixed-combination, 272–273
- Drug Price Competition and Patient Term Restoration Act, 3, 29, 116
- Duration, maximum, 420, 428, 430
- Eastern Cooperative Oncology Group (ECOG), 202
- Effect, 73, 77
 adverse, 96
 block, 142
 carryover, 58–59, 179–181, 187, 436
 center, 534
 common treatment, 528
 confounding, 44, 55, 57
 direct drug, 180
 drift, 55
 drug, 180
 estimated treatment, 416, 523
 fixed, 326, 475
 direct, 475
 interaction, 44, 55
 main, 60
 optimal therapeutic, 117
 overall average drug, 326, 331
 overall treatment, 533
 overall trend, 514
 period, 436
 pharmacokinetic, 14
 pharmacological, 14
 placebo, 102, 107, 513–516
 prophylactic, 386
 random, 331, 474
 residual, 168, 179
 sequence, 58, 181, 436
 sensitivity-to-drug, 78, 81, 255–256
 side, 14, 20, 549
 subject, 566
 synergistic, 271
 therapeutic, 117, 168, 271
 time, 327, 331, 514
 treatment, 19, 53, 57–59, 128, 179, 436
 consistent, 533
 overall, 533–534
 treatment-by-time, 326
 trend, 514
- Effectiveness, 1, 2, 6, 8, 22, 31, 64, 75, 79, 91, 103, 110, 261
- Efficacy, 1–3, 6–7, 17, 24, 31, 47, 55, 71, 77, 90, 104, 290
 average, 155
 clinically meaningful, 168
 composite, 294
 false negative, 59
 false positive, 59
 individual, 155–158
 average, 158
 noninferior, 240
 population, 155–156
 relative vaccine, 294
 superior, 106–107, 542, 545
- Electrophysiologic Study versus Electrocardiographic Monitoring (ESVEM) Trial, 196–199
- EM algorithm, 189, 559

- Endpoint, 6, 14, 89, 307, 318
 binary, 353, 385, 481
 independent, 482
 paired, 487
 binary clinical, 360, 560
 categorical, 339–340, 386, 529
 censored, 386
 independent, 489
 censored clinical, 435
 paired, 436
 clinical, 54, 66, 109, 386
 efficacy, 512
 safety, 512
 clinical surrogate, 45
 continuous, 365, 386, 528
 efficacy, 17, 60, 64, 66, 89
 ordinal categorical, 341
 exploratory, 541
 hard, 159
 multiple, 511–512, 539–543
 multiple binary clinical, 543
 multiple efficacy, 538
 normal continuous, 532
 primary, 48, 423, 486, 538, 542
 primary clinical, 47, 55, 66, 94, 126, 339, 437, 512
 primary efficacy, 17, 19, 91, 93, 247, 257, 387, 434
 primary safety, 92
 repeated categorical, 341
 safety, 17, 19
 secondary, 541–542
 secondary efficacy, 91, 93, 437
 superior, 247
 surrogate, 27, 103
 survival, 481
 tertiary, 541, 542
- Entry, staggering, 420
- Equality, 81, 474, 493
- Equation
 Fieldewald, 569
 generalized estimating (GEE), 189, 326, 331, 336–337, 529
 maximum likelihood, 484
 normal, 373
- Equivalence, 81, 476, 483, 579
 average, 298
 clinical, 44, 77, 111
 one-sided, 385
 population, 298
 therapeutic, 106, 384, 440
 two-sided, 252, 257
- Equivalence limit, 481
- Equivalence/noninferiority, 81, 90
- Erectile dysfunction, 90, 106, 193, 206
- Error
 maximum, 442
 measurement, 54–55
- random, 48, 53
 within-subject, 420
- systematic, 420
 type I, 71–72, 112, 122, 417, 419, 427, 438, 441
 type II, 71, 438, 441
 within-subject, 420
- Establishment License Application (ELA), 13, 15
- Estimate, 65, 336
 Bayesian, 224
 best linear unbiased, 514
 consistent, 128, 236, 334, 373
 estimated generalized least squares, 516
 inconsistent, 333
 inefficient, 333
 interval, 60, 70, 301
 Kaplan-Meier, 392–394, 414
 Kaplan-Meier survival, 393, 397, 411
 least squares (LSE), 526
 maximum likelihood (MLE), 335, 373, 406–407, 470, 560
 point, 66, 68, 301, 566
 unbiased, 470
- Estimation, 301
 interval, 343
 large sample, 402
 Kaplan-Meier nonparametric, 392
 sample size, 89, 109, 110
 unbiased, 437
- Estimator
 best linear unbiased, 514
 combined, 535
 consistent, 334, 536
 estimated generalized least squares (EGLS), 515
 least squares, 318
 maximum likelihood, 373, 380, 488
 modified, 437
 restricted, 484
 median unbiased, 437
 minimum variance unbiased (MVUE), 535–536, 546
 point, 437
 quasi-likelihood, 334
 unbiased, 58, 501, 526, 533
 unbiased consistent, 515
- Estrogen Replacement and Atherosclerosis (ERA) trial, 556
- European Community (EC), 24
- European Federation of Pharmaceutical Industries Associations (EFPIA), 35
- European Organization for Research and Treatment of Cancer (EROTC), 216, 574
- Evaluation, 1, 2, 17
 image, 209
 blinded, 209
 combined, 211
 consensus, 210
 fully blinded, 209

- independent, 210
- onsite, 210
- offsite, 210
- primary, 209
- separate, 210
- unblinded, 209
- unpaired, 210
- laboratory, 608
- safety, 77, 110
- Evidence
 - qualitative, 437
 - quantitative, 437
 - substantial, 8, 22, 75, 84
- Expanded access, 24
- Expectation, 333
 - marginal, 334
- Expected value, 323, 526
- Experimental unit, 1, 2, 17, 169, 174
- Factor, 275
 - classification, 534
 - confounding, 59, 110, 523
 - demographic, 118
 - designed, 534
 - ethnic, 285
 - extrinsic, 285–286
 - intrinsic, 285
 - expected bias, 128, 135
 - finite population correction, 365
 - fixed, 533
 - prognostic, 130, 144, 414, 509, 523
 - random, 533
 - risk, 523
 - variance inflation (VIF), 175
- Federal Food, Drug, and Cosmetic Act (FD&C), 3, 6, 8, 11, 562
- Federal Register, 34, 168, 564
- Fibonacci sequence, 217
- Fieller theorem, 236
- Food and Drug Administration (FDA), 3, 8, 32, 100
 - inspection, 613
 - organization, 8–12
- Food and Drug Administration Modernization Act (FDAMA), 4, 12
- Forced expiratory volume in one second (FEV1), 91
- Fred Hutchinson Cancer Research Center (FHCRC), 176
- Function
 - baseline hazard, 403–404, 409
 - cumulative distribution, 389, 479
 - cumulative standard normal distribution, 470
 - hazard, 391, 403, 406
 - proportional, 403
 - Kaplan-Meier survival, 393, 397
 - identity, 413
 - likelihood, 224, 415
 - link, 334
- logistic, 217
- log-likelihood, 409–410, 414, 560
- monotone, 431
- power, 439, 472
- probability density, 391
- quadratic error loss, 224
- quasi-score, 333
- spending, 431
 - alpha, 388, 423, 427, 430, 433
 - continuous, 431
- step, 408
- survival, 234, 338, 387, 390, 394, 489
- General search category, 575
- Generalizability, 82, 84, 239, 243, 284, 315
- Generic drug product, 29, 116
- Genetics, 510–511, 552
- Genotype
 - inert, 557
 - reactive, 557
- Gold standard, 268
- Good clinical practice (GCP), 32–35, 93, 109, 602, 604, 617, 632
- Good data management practice (GDMP), 629–630
- Good laboratory practice (GLP), 629–630
- Good programming practice (GPP), 645
- Good statistical practice (GSP), 35, 630
- Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries Trial (GUSTO I or Global Use of Strategies to Open Occluded Coronary Arteries), 7, 45, 52, 153, 162, 262
- GUSTO III, 262
- Greenwood formula, 393
- Growth curve, 337
- Hamilton depression scale, 78, 79, 104
- Hardy-Weinberg equilibrium, 536
- Health maintenance organization (HMO), 4
- Health status survey short form, 36, 248
- Helmert matrix, 517
- Helmert transformation, 514
- Hematology, 150, 608
- Heterogeneity, 536
- High level group term (HLGT), 581
- High level term (HLT), 581
- Holter monitor, 45, 198–199
- Holter recording, 190
- Homogeneity, 531
- Homoscedasticity, 537
- Hormone Replacement Therapy (HRT), 433, 556
- Hotelling T^2 , 514–555
- Human, 1, 2
- Human immunodeficiency virus (HIV), 27, 97, 103, 107, 256
- Hypercholesterolemia, 6

- Hypothesis
 alternative, 19, 71–73, 305, 394, 544
 specific, 444
 equivalence, 489
 interval, 80, 112, 252, 384, 476, 479
 noninferiority, 261, 486, 491
 null, 19, 71–77, 80, 124, 306, 309, 318–319
 ordered, 540
 one-sided, 75, 80, 247–248, 306, 476, 489
 point, 80, 112, 474, 479
 primary, 607
 secondary, 607
 statistical, 17, 19, 394, 609
 superiority, 261
 tertiary, 607
 two one-sided, 273, 477
 two-sided, 75, 247, 305, 444
 two-sided alternative, 559
- Human genome project (HGP), 553
- Hutchinson Smoking Prevention Project (HSPP), 176–177
- Hypothesis of equality, 476
- Hypothesis testing, 43–44, 65–66, 71, 306
- Imbalance, 141
 covariate, 127–128, 136–138, 145
 treatment, 133–136
- Immune response, 292
- Immunogenicity, 289–290
- Incidence, disease, 292
- Index
 Chalmer's, 622
 Jadad's, 622
 sensitivity, 86–87
- Inference
 biased, 518
 clinical, 89–90, 93, 101, 104, 109
 statistical, 17, 23, 44–47, 90–93, 101, 109, 114, 122, 301, 344
 unbiased, 517
 unbiased, 89
- Information, 419, 420
 clinical, 419, 628
 genetic, 552–553
 genomic, 511
 individual, 428
 maximum, 420–421, 432
 statistical, 419
 total, 420, 430
- Information matrix, 373, 407
- Informed consent form, 100, 103
- Inspection, 612, 617
- Institutional review board (IRB), 20, 21, 100, 602, 608, 618
- Interaction, 43, 60–64, 297
 crossover, 64
 drug-to-drug, 168, 282, 562
- qualitative, 63–64, 242, 246, 536
 quantitative, 63–64, 242, 245, 536
 time-covariate, 409
 treatment-by-age, 415–416
 treatment-by-center, 62, 116, 241, 244–245, 315–316, 535–536
 qualitative, 241–242, 246, 531
 quantitative, 241–242, 246, 531
 treatment-by-covariate, 546
 treatment-by-effect, 537–538
 treatment-by-period, 58
 treatment-by-study, 247
 treatment-by-time, 44, 329, 415–416
 treatment-by-visit, 331
 two-factor, 375
- Interaction between treatment and time, 468
- Interaction mean square, 534
- Interactive Voice Randomization System (IVRS), 646–647
- International Classification of Diseases (ICD), 575
- International Conference on Harmonisation (ICH), 34–35, 511
- International Index of Erectile Function (IIEF), 194
- International Federation of Pharmaceutical Manufacturers Association (IFPMA), 35
- Intervention, 2, 433
 therapeutic, 127, 174
- Investigational Device Exemptions (IDE), 13
- Investigational New Drug Application (IND), 13–16, 21, 631
 commercial, 15
 noncommercial, 15
 rewrite, 14
 treatment, 6, 21
- Investigator, 23, 159
 clinical, 602
 principle, 604
- In Vitro*, 168, 200, 257
- In Vivo*, 168
- Iterative reweighted least squares, 373
- Japanese Pharmaceutical Manufacturers Association (JPMA), 35
- Kefauver—Harris Amendment, 3, 562
- Kronecker product, 590
- Last observation carried forward (LOCF), 337, 519–520
- Latin squares, orthogonal, 183
- Laboratory
 central, 550, 592, 608
 local, 592
- LD10, 217
- Level, 275
 confidence, 66, 77, 441
 nominal, 345, 424

- significance, 72, 110, 122, 418, 430, 609
 - nominal, 424, 542, 544, 546
 - overall, 424, 431
- Level of significance, 71–77, 306, 423, 441
 - nominal, 19, 89
- Limit
 - equivalence, 81, 255, 259
 - equivalence/noninferiority, 259–261
 - noninferiority, 260–261
 - similarity, 496
- Lilly reference, 593
- Link
 - complementary log, 374
 - identity, 374
 - log, 374
 - logit, 374
 - probit, 374
- Lipid Research Clinics Coronary Primary Prevention Trial (CPPT), 154
- Location, 46
- Location shift, 322–323, 365, 370
- Logit, 372
 - cumulative, 377
- Lost to follow-up, 386, 417
- Lowest level term (LLT), 580
- Masking, 113, 158
- Maximum urinary flow rate, 102
- Mean, 65, 301
 - arithmetic, 46
 - marginal, 334
 - overall sample, 311, 327
 - population, 301, 419, 432, 560
 - sample, 66, 309, 318
- Mean square error (MSE), 312, 331, 437
- Measure, repeated, 301, 326, 463
- Measurement, repeated, 333, 337, 538
 - baseline, 512–514
 - multiple, 513
- Median, 46
- Medical device, 3, 8, 11
- Medical Dictionary for Drug Regulatory Affairs (MedDRA), 563, 577
- Medical monitor, 520
- Medical Research Council, 4
- Medication event monitor system (MEMS), 117
- Medicine, 1
- Median survival, 235
- Metered dose inhaler (MDI), 106
- Method
 - continual reassessment, 41, 216, 223, 235–236
 - delta, 236
 - estimated generalized least squares (EGLS), 543
 - exact, 361
 - GEE, 334
 - group sequential, 297, 388, 422, 437
 - Haybittle and Peto, 425
- Kaplan-Meier, 387
- least squares, 282
- likelihood ratio, 236
- life-table, 189
- log-likelihood, 374–375
- logrank, 397
- maximum likelihood, 282
- minimization, 144
- model-based, 343, 372
- nonparametric, 392
- parametric, 392
- randomization-based, 343
- response surface, 283
- scoring, 374–375
- stochastic curtailment, 432
- two-stage active control testing (TACT), 264
- Missing value, 89, 117, 560, 609
 - completely random, 561
 - informative, 561
 - intermittent, 338
 - random, 561
- Missing at random (MAR), 118, 385, 559
- Missing completely at random (MCAR), 385, 561
- Ministry of Health, Labor, and Welfare (MHLW) of Japan, 32
- Mode, 46
- Model
 - analysis of covariance, 523
 - analysis of variance, 452, 476, 532
 - one-way, 526
 - two-way, 532
 - Anderson-Gill, 566
 - autoregressive, 337
 - cell-means, 535
 - Cox's proportional hazard, 402–410, 529, 546
 - stratified, 409–410
 - Cox's proportional hazard regression, 403, 405
 - fixed effect, 454
 - frailty, 566
 - generalized linear (GLM), 323, 454
 - hierarchical, 288
 - interaction, 408
 - invoked population, 123
 - linear, 128
 - linear random effect, 419
 - linear regression, 224, 463
 - logistic linear dose-response, 189
 - loglinear, 384
 - main-effect, 534
 - marginal, 333–337
 - Markov, 466
 - mixed effect, 534
 - nested, 326
 - nonlinear, 128
 - permutation, 142
 - population, 112
 - population average, 337

- proportional hazard, 234, 388
- proportional hazard regression, 403
- proportional odds, 377–378
- random effects, 337, 566
- randomization, 122–123
- randomized complete block, 454
- reduced, 409
- subject-specific, 337
- transition, 333
- two-way classification fixed, 316–317
- Monitoring**, 617
- Monotherapy**, 275, 279, 297
- Mortality**, 60–62, 115, 241
- Multicenter Automatic Defibrillator Implantation Trial II**, 201
- Multiple baseline measurement**, 513
- Multiple-evaluator**, 162
- Multiple-placebo**, 162
- Multiplicity**, 537–538, 609
- Myocardial infarction**, 45, 60, 92–93, 115, 153, 201, 260, 339
- N-of-I randomized trial**, 157
- National Cancer Institute (NCI)**, 4, 216
- NCI Common Toxicity Criteria (CTC)**, 216, 574, 581
- National formulary**, 12, 29
- National Heart, Lung, and Blood Institute**, 45, 198, 418, 433
- National Institute of Allergy and Infectious Disease**, 552
- National Institutes of Health (NIH)**, 3, 417, 433
- National Institutes of Health Reauthorization Bill**, 154
- National Institutes of Health Stroke Scale (NIHSS)**, 55–63, 126, 340, 607
- National Institute of Neurological Disorder and Stroke (NINDS)**, 55, 340
- National Surgical Adjuvant Breast and Bowel Project (NSABP)**, 614
- Negative event**, 571
- New Drug Application (NDA)**, 6, 13, 15, 22, 24, 631
 - abbreviated (ANDA), 13, 29, 631
 - supplemental (SNDA), 30
- Newton—Raphson algorithm**, 373
- Noninferiority**, 81, 493
 - clinical, 81
- Noninsulin-dependent diabetes mellitus (NIDDM)**, 95, 172, 186
- Nonparametrics**, 320
- Normality**, 371
- Numerical integration**, 428
- Objective**
 - primary, 90–92, 251
 - primary study, 17, 439
 - secondary, 17, 90–92
 - tertiary, 83
- Open-label**, 121, 159–160
- Outcome**
 - binary, 486
 - paired, 487
- Outcomes research**, 2
- Parallel track regulation**, 6, 27
- Parameter**, 46–47, 68, 83
 - efficacy
 - primary, 451, 607
 - secondary, 607
 - tertiary, 607
 - noncentrality, 446, 453
 - nuisance, 333, 463
 - overdispersion, 334
 - safety, 608–609
 - scaled, 333
- Part 11 compliance**, 630–631
- Partial likelihood**, 406, 410
- Patient**, 1, 2, 159, 326
 - evaluable, 508
 - qualified, 508
- Patient characteristic**, 512
- Patient data listing**, 511
- Patient identification number**, 113
- Patient identifier**, 522
- Peak urinary flow rate**, 63, 123, 125
- Period**
 - accrual, 200
 - active treatment, 117, 171
 - baseline, 512
 - follow-up, 420, 604
 - lead-in, 171
 - placebo run-in, 187, 435, 513
 - recruiting, 422
 - run-in, 17, 171, 256
 - single-blind placebo run-in, 173
 - treatment, 179, 187
 - washout, 58, 171, 179, 187, 256
- Permutation**, 125
 - random, 135–139
- Pharmaceutical entity**, 339
- Pharmaceutical identity**, 1, 31
- Pharmaceutical Research and Manufacturers of America (PhRMA)**, 35
- Pharmacoeconomics**, 2
- Pharmacogenomics**, 2, 7
- Pharmacokinetics/pharmacodynamics (PK/PD)**, 286, 622
- Phase**
 - active, 572
 - dose-titration enrichment, 194
 - double-blind treatment, 173
 - enrichment, 194, 198, 611
 - maintenance, 117
 - placebo baseline, 1, 94
 - double-blind, placebo-controlled, 194

- placebo run-in, 117
- placebo washout, 189
- run-in, 611
- run-in single-blind, 173
- sustained active, 184
- Phase I clinical investigation, 14
- Placebo, 2, 52, 60, 90–91, 102, 120, 124, 127
 - matching, 423
- Placebo concurrent group, 103
- Placebo responder, 196
- Play-the-winner role, 145
 - modified (MPW), 145–146
 - randomized (RPW), 145–147
- Population, 2, 71
 - efficacy patient, 520
 - geriatric, 100
 - homogeneous, 122, 125
 - intention-to-treat, 538
 - patient, 46–47, 114, 122
 - pre-protocol patient, 520
 - targeted, 2, 17, 46, 52, 65, 83, 333, 512
 - targeted patient, 43, 53, 84, 89, 93–94, 114, 122, 215, 243, 301, 602, 611
 - time-heterogeneous, 123, 140
- Posterior density, 224
- Postrandomization discontinuation, 522
- Postmenopausal Estrogen/Progestin Interventions Trial (PEPI), 417
- Power, 72–73, 97, 110–111, 122, 133, 141, 438, 441, 541, 609
 - conditional, 432
 - estimated, 83
- Precision, 47, 53–54, 89, 98, 438
 - optimal, 126
- Premarket Approval of Medical Devices (PMA), 13, 15
- Premature discontinuation, 519
- Premature withdrawal, 520
- Premature ventricular beat (PVB), 45
- Prescription Drug User Fee Act (PDUFA), 4
- Price Competition and Patent Restoration Act, 3
- Primary pulmonary hypertension, 108
- Principle
 - intention-to-treat, 517–520, 560
 - uncertainty, 626
- Probability, 43–46, 128, 406
 - conditional, 83, 391, 406
 - coverage, 345
 - discordant, 487
 - generalizability, 84–87, 284, 285
 - posterior, 224
 - reproducibility, 83, 284
- Procedure
 - blinding, 510
 - Bonferroni, 530–540
 - Dunnett, 541–542
 - exact test, 351
 - generalized least square (GLS), 516
- group sequential, 423–424
 - K-stage, 424
 - O'Brien-Fleming, 428–430, 437
 - Pocock, 425, 430
- group sequential bioequivalence testing, 437
- Hochberg, 540
- Holm's Bonferroni, 540–541
- intersection union test (IUT), 545
- nonparametric, 544
- one-step, 248
- semiparametric statistical, 403
- two one-sided tests, 477, 490
- unconditional exact, 356
- two-stage, 247
- unconditional large-sample, 355
- Process
 - random, 535
 - titration, 192
- Program to Assess Alternative Treatment Strategies to Achieve Cardiac Health (PATCH), 104, 248
- Product License Application (PLA), 6, 13, 15
- Programmer, 149, 159–160
- Proportion, 357
 - binomial, 345
 - marginal, 381
 - population, 344, 356
 - proportional hazard, 407
 - sample, 344, 444
- Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial, 248, 609
- Protocol, 17–18, 43–44, 112, 114, 122
 - amendment, 20, 613
 - clinical, 602
 - component, 603
 - deviation, 603, 612
 - fraud, 603–604
 - implementation, 603
 - misconduct, 603, 614
 - preparation, 603
 - structure, 603
 - violation, 520, 603, 613
- Protocol review committee, 33, 149
 - internal, 149
- Pure Food and Drug Act, 3
- P-value, 43–44, 71–73, 82, 112, 124–125, 246, 306
 - conditional exact, 335
 - observed, 74–75, 83
- Quality assurance, 150, 601
- Quality of life, 2, 7, 31, 66–68, 92, 104, 326
- Query form, 639
- Random allocation, 129, 134–135, 139
- Random assignment, 122–129, 139, 150–153
- Random censoring, 392
- Randomization, 4, 8–9, 19, 23, 117, 120–123, 127, 165–166, 239

- adaptive, 121, 129, 142
 - covariate, 129, 142, 144
 - response, 129, 142, 145
 - treatment, 129, 142
- bias coin, 142, 143
- complete, 121, 129–136, 142
 - cluster, 176
 - minimization, 153
 - permuted-block, 121, 129, 136–139, 142
 - restricted, 132, 142
 - simple, 131–137
 - stratified, 126–127, 141, 290, 340
 - unstratified, 127
- urn, 143–144
- within-trial, 263
- Randomization code, 113, 129, 131–135, 149, 298, 519
- Randomization method, 127–128
- Random number, 131
- Random selection, 122, 129, 139
- Range, 46
 - equivalence, 575
 - normal, 386, 593–594
 - reference, 592
 - referenced laboratory, 362
 - standard, 592
 - therapeutic dose, 282
- Rank, 321
- Rapid eye movement latency (REML), 147
- Rate, 565
 - crude incidence, 564–565
 - cumulative event, 466
 - dropout, 337, 519
 - eradication, 272
 - experimentwise false positive, 539, 544, 546
 - experimentwise type I error, 539, 544
 - event, 432
 - false negative, 541
 - false positive, 430, 541–542
 - overall, 542
 - cumulative error, 429
 - hazard, 464
 - incidence, 451, 567
 - instantaneous death, 391
 - maximum uroflow, 537
 - mortality, 390
 - prevalence, 567
 - relative hazard, 398, 402
 - response, 226, 234, 558
 - survival, 555, 565
 - type I error, 417, 438, 492, 520
 - overall, 509, 639
- Ratio
 - benefit-to-risk, 280
 - hazard, 234, 405–408, 436
 - odds, 357
 - cumulative, 357
- relative hazard, 402
- signal-to-noise, 84
- Recurring event, 567–568
- Region
 - acceptance, 74, 306
 - rejection, 71, 74, 306
- Regression
 - linear, 378
 - logistic, 372, 529, 546, 588
- Regression coefficient, 333, 373, 410, 526
 - estimated, 335
 - individual, 436
- Regulation, 13
- Relationship, dose-response, 58, 88, 282, 468, 540
- Reliability, 43, 47, 53, 89, 93
- Repeatability, 593
- Report,
 - clinical, 510, 523
 - full integrated, 510
 - integrated clinical, 511
 - integrated summary safety, 531
 - integrated statistical, 511
 - safety, 21
 - statistical, 510
- Reproducibility, 51, 82, 243, 246, 248, 315, 593
- Response
 - binary, 340, 353
 - immune, 292
 - therapeutic, 436
- Response Evaluation Criteria in Solid Tumor (RECIST), 216, 607
- Response surface, 281
 - estimated, 282
- Risk
 - consumer's, 72
 - exposure, 565
 - false positive, 428
 - producer's, 72
 - relative, 357–358, 404–407, 451, 486,
- Risk set, 406
- Safety, 2, 3, 6–8, 14, 17, 22, 24, 31, 47, 55, 66, 71, 77, 89–90, 103–104, 110, 118, 290
- Safety assessment, 109
- Safety Medical Device Act (SMDA), 13
- Sample
 - paired, 385
 - random, 71, 104, 301
 - representative, 46, 53, 114, 122, 333
- Sample size, 47, 53, 77, 111, 114, 125–133, 297, 423, 438
 - adjustment, 560
 - determination, 19, 111–112, 438, 443, 609
 - estimation
 - censored data, 464
 - crossover design, 474
 - dose response, 468

- equivalence and noninferiority, 481
- multiple samples
 - ANOVA, 452
 - GLM, 454
- multiple-stage design, 492
- one sample, 443
- two sample, 445
- variability
 - intersubject, 501
 - intrasubject, 494
- expected, 226–228, 438
- justification, 111
- planned, 559
- reestimation, 558–560, 609
- Sampling without replacement**, 135
- Scale**
 - nominal, 339–340
 - ordinal, 339
- Score**
 - American Urinary Association (AUA) symptom, 63
 - integer, 365
 - Karnofsky performance, 387
 - modified Rankin, 93, 365
 - logrank, 365
 - positive and negative symptom (PANSS), 84
 - standardized midrank, 365
- Second International Study of Infarct Survival (ISIS2)**, 60–61, 153, 273, 296
- Sequential equivalence testing**, 284
- Session**
 - closed, 551
 - executive, 551
 - open, 551
- Sensitivity**
 - assay, 78, 169, 255
 - ethnic, 284
 - global, 287
- Sherley Amendment**, 2
- Single nucleotide polymorphism (SNP)**, 536
- Significance**
 - clinical, 77, 510
 - comparative clinical, 78–79
 - individual clinical, 78
 - prognostic, 512
 - statistical, 78–79
- Similarity**, 86–87, 175, 287, 483, 486
 - population, 287
- Similarity circle**, 156–157
- Size**, 544
 - block, 140–141, 166
 - effect, 86, 255
 - effective, 260, 263
- Skewness**, 46
- Southwest Oncology Group (SWOG)**, 27
- Special search category**, 575, 581
- Sponsor**, 159
- Spontaneous events**, 572
- Spread**, 46
- Standard**
 - gold, 208
 - true, 208
- Standard deviation**, 46, 65, 301–302, 309, 442
 - population, 302
 - sample, 302
 - within-group sample, 543
- Standard operating procedure (SOP)**, 32, 121, 149, 618
- Standard error**, 47, 65–66, 70, 302, 336, 423
 - large-sample, 406, 414
- Statistical analysis plan (SAP)**, 164
- Statistics**, 4
 - Bhapkar Q, 382
 - descriptive, 43–44, 65–66, 532
 - Gehan test, 402
 - inferential, 43, 65–66
 - likelihood ratio, 463
 - linear rank, 142
 - logrank, 144, 398–399, 465
 - logrank test, 397–398, 501
 - Mantel—Haenszel, 367–368, 397–398
 - extended, 370
 - McNemar, 599
 - minimax, 246
 - paired Prentice-Wilcoxon, 436
 - Peto-Prentice-Wilcoxon, 144
 - sample, 47, 65
 - score, 503
 - Stuart-Maxwell, 597
 - summary, 549
 - t-, 303
 - two-sample, 475
 - two-sample unpaired, 546
 - test, 82, 305, 351
 - asymptotic, 355
 - large sample randomization, 365
 - Wald, 407
 - Wilcoxon-Mann-Whitney test, 322
 - Wilcoxon rank sum, 322
 - blocked, 370
 - Z-, 423
 - Steering committee, 550, 551
 - Stratification**, 23, 54, 121, 126
 - Stratum**, 141
 - Streptokinase**, 45, 60, 62
 - Study**
 - active control, 511–512
 - ADME, 32
 - adequate and well-controlled, 22–23, 71, 101, 153, 260
 - adequate and well-controlled clinical, 71
 - bridging, 86, 239, 283–289, 291
 - clinical, 87
 - blind reader, 152, 155
 - case-control, 126
 - CNAAB3005 International, 257

- dose-escalating, 160, 189
- dose proportionality, 105
- dose response (ranging), 16, 296, 468
- dose titration, 54, 57
- double-blind, three-arm, parallel, randomized, 313
- Helsinki Heart (Health), 2, 7
- immunogenicity
 - dose-response, 291
 - superiority, 291
- immunogenicity persistent, 292
- multivalent vaccine, 292
- multicenter, 19, 116
- National Institute of Neurological Disorder and Stroke rt-PA Stroke (NINDS rt-PA Stroke Study), 62, 126, 153, 344
- observational, 166
- parallel-group, 173
- pharmacokinetic/pharmacodynamic, 286
- phase I, 14
- phase I and II, 16
- phase I safety and tolerance, 188
- phase II, 14
- phase IIA, 14
- phase IIB, 14
- phase II and III, 14, 547
- phase III, 14, 543
- phase IIIB, 14
- phase III and IV, 113
- phase IV, 15, 569
- phase V, 15
- Physicians' Health, 2, 7, 154, 339, 358
- placebo-controlled, double-blind, 386
- positive control, 112
- postmarketing surveillance, 160
- premarketing surveillance, 160
- prospective cohort, 446
- randomized, 154, 386
 - multicenter, parallel-group, 323
- retrospective case-control, 446
- single-site, 241
- Systolic hypertension in the elderly (SHEP), 7
- titration, 58
- two-group parallel, 386
- uncontrolled, 100
- Study center, 98, 159
- Study site, 98
- Subgroup, 564
- Subject, 1, 2, 3
- Subpopulation
 - heterogeneous, 123
 - homogeneous, 123
- Sum of squares, 133
 - correct, 420
 - due to center, 316
 - due to error, 312–314, 319
 - due to treatment, 311–319, 337
 - error, 327
- residual, 527
- total, 311, 316
- Summary statistical table, 511
- Superiority, 493
 - global, 287
 - strict, 279–281, 296
 - wide, 280–281
- Survival
 - event-free, 403
 - exponential, 464
 - median, 235, 386
 - overall, 403
- Symmetry, compound, 436
- System organ class (SOC), 580
- Term
 - high level, (HLT), 580
 - high level group (HLGT), 580
 - lowest level (LLT), 579
 - preferred, (PT), 579
- Test
 - blocked Wilcoxon rank sum, 370, 528
 - Breslow-Day, 591
 - chi-square, 335, 360–361
 - randomized, 361
 - Cochran-Mantel-Haenszel, 590–591
 - distribution-free, 321
 - F-, 213–319, 327, 452
 - Fisher exact, 361, 586, 589
 - Gehan, 399
 - goodness-of-fit, 282
 - intersection-union, 544
 - Kolmogorov-Smirnov, 299
 - Kruskal-Wallis, 125, 320, 325–326, 337, 365
 - lack-of-fit, 282
 - linear rank, 125
 - logrank, 394, 399–401, 405, 433, 465
 - log-likelihood chi-square, 373
 - Mann-Whitney, 322
 - Mantel-Haenszel, 142, 367, 586–599
 - McNemar, 355, 365, 488, 600
 - one-sample, 380
 - two-sample, 380, 381
 - Min, 280
 - Newman-Keuls range, 327
 - nonparametric, 399
 - one-sample, 443
 - one-sided, 75, 76
 - paired t , 301, 306–307
 - Pearson, 373
 - Pearson chi-square, 360–361
 - permutation, 123–125, 431
 - conditional, 125
 - unconditional, 125
 - Peto-Peto-Prentice-Wilcoxon, 125
 - randomization, 144
 - randomized chi-square, 361

- score, 407
- sign, 436
- statistical, 73, 75, 79
- Stuart-Maxwell, 597
- Student *t*-, 297
- t*-, 337
- treadmill, 435
- two one-sided, 479
- two-sample, 445
- two-sample *t*-, 77, 301–308
- two-sample Z-, 443
- two-sided, 73–84
- Wilcoxon-Mann-Whitney, 322
- Wilcoxon rank sum, 125, 297, 321, 337, 365, 543
- Wilcoxon signed rank, 320–323
- Wilcoxon two-sample rank sum, 399
- Williams, 470–473
- Therapeutic range, 168, 188
- Therapeutic window, 188
- Time, 329
 - calendar, 388, 417–421, 427
 - censoring, 391, 464
 - discrete information, 421
 - failure, 391
 - information, 417–421, 427
 - median failure, 436
 - median survival, 386, 393, 404
 - survival, 391–392
- Titration
 - crossover, 17
 - dose, 326
 - forced, 17
 - parallel, 17
- T-PA, 45, 63
- Tolerability, 91
- Toxicity, 92, 564, 568
 - dose limiting (DLT), 216–220
 - systematic, 563
- Transcutaneous electrical nerve stimulation (TENS), 163
- Treatment, 1, 2, 17
 - active, 100, 250
 - actually given, 517
 - planned, 517
 - planned randomized, 518
- Treatment-emergent events, 600
- Treatment group, 276
- Trial
 - active control, 77, 107, 251, 253, 256
 - two-arm, 25, 254, 255, 263
 - active control equivalence (ACET), 252–256, 260
 - adequate, well-controlled clinical, 11, 30, 43, 75, 82, 100, 283, 510
 - bioequivalence, 70, 115, 240
 - clinical, 1–4, 17, 422
 - quality assessment, 620
 - combination, 239–240, 270, 282
 - comparative, 159
 - comparative clinical, 120
 - confirmatory clinical, 112
 - Continuous Infusion versus Double-bolus Administration of Alteplase (COBALT), 107, 259
 - controlled clinical (CCT), 286
 - controlled randomized, 114, 127, 154
 - dose-ranging, 52
 - dose-response, 239–240, 265
 - double-blind, placebo-controlled, 172
 - double-blind, randomized multicenter clinical, 537
 - double-blind, randomized, placebo-controlled, 104
 - equivalence, 239, 250, 256, 481
 - active control, 252
 - two-sided, 251–252, 481
 - equivalence/noninferiority, 239, 250
 - extracorporeal membrane oxygenation (ECMO), 146–147
 - Lipids Research Clinic Coronary Primary Prevention, 154
 - maximum duration, 421
 - maximum information, 421
 - multi-arm, 229
 - multicenter, 98, 114, 141, 239–240, 244, 511, 529–530, 608
 - multicenter clinical, 242
 - noninferiority, 92, 107, 250, 252, 481
 - one-sided, 252, 481
 - phase I, 6, 289
 - phase I cancer, 216, 236
 - phase I clinical, 16, 27
 - phase II, 113, 289, 542
 - phase II cancer, 216, 238
 - phase II clinical, 27, 188–189
 - phase III, 16, 289
 - phase III clinical, 27
 - phase III confirmatory, 387
 - phase IV, 15
 - phase I, II, and III, 4, 24
 - pivotal, 82, 283
 - placebo-controlled, 256
 - placebo-controlled clinical, 508
 - primary prevention, 2, 433
 - Prostate Cancer Prevention (PCPT), 7
 - randomized, controlled, 165
 - randomized, double-blind, parallel-group clinical, 546
 - randomized, double-blind, placebo-controlled, 418
 - randomized, triple-blind, two-parallel-group, 423
 - sequential, 239, 298
 - single-center, 114
 - single-site, 529
 - superiority, 75, 239, 247, 251, 263, 481
 - telescoping, 27
 - vaccine, 240, 289
 - vaccine clinical, 139, 289

- well-controlled randomized clinical, 154
- women-health, 2
- Trial monitor, 617–618
- Unbiasedness, 44, 47
- Uncertainty, 43–44
- Uncertainty principle, 627
- Urinalysis, 150, 608
- U.S. master drug files, 285
- U.S. National Cancer Institute, *see* National Cancer Institute
- U.S. National Heart, Lung, and Blood Institute, *see* National Heart, Lung, and Blood Institute
- U.S. National Institute of Allergy and Infectious Disease, *see* National Institute of Allergy and Infectious Disease
- U.S. National Institutes of Health (NIH), *see* National Institutes of Health
- U.S. National Institute of Neurological Disorder and Stroke, *see* National Institute of Neurological Disorder and Stroke
- U.S. Pharmacopeia, 12, 29
- Validation, external, 255, 275
- Validity
 - assay, 90
 - external, 114–115
 - internal, 114, 620–622
- Variability, 43, 47, 53, 65, 79, 89, 94, 97, 110, 155, 301, 439
 - between-patient, 329
 - between-site, 593
 - interpatient, 53, 170, 179, 212
 - intersubject, 493
 - intrapatient, 158, 170, 212
 - intrasubject, 54, 182, 493
 - within-laboratory, 593
- Variability of efficacy, 155
- Variable
 - baseline, 512
 - binomial, 356
 - continuous, 301
 - derived, 636
 - discrete, 301
 - explanatory, 333–334, 375
 - latent, 362
 - Poisson, 374
 - primary efficacy, 109
 - primary response, 109
- primary safety, 110
- prognostic, 512
- random, 302
 - Bernoulli, 444
 - binomial, 345
 - continuous, 389
 - chi-square, 361, 368, 591, 596. *See also* Variable, random, χ^2
 - F, 494
 - normal, 465
 - χ^2 , 355
- response, 519
 - secondary response, 109
 - standard normal, 446
 - time-dependent, 403
- Variance, 46, 323, 437, 532
- estimated, 542
- estimated large sample, 357–359
- intersubject, 501
- intrasubject, 498
- large-sample, 380
- marginal, 334
- pooled, 534
- population, 301, 417, 446
- sample, 309
- within-group, 559
- Variation, 46, 53, 120
 - biological, 45, 53, 156
 - intersubject, 53–54
 - intrasubject, 54
 - random, 311
 - temporal, 53–54
- Ventricular ectopy, 190
- Ventricular premature contraction (VPC), 103, 190, 196
- Ventricular tachycardia, 538
- Veteran Administration Cooperative Urological Research Group, 123
- Weighted least squares, 383
- West of Scotland Coronary Prevention Study Group, 452
- Women's Health Initiative (WHI), 433, 555, 609
- World Health Assembly (WHA), 106
- World Health Organization (WHO), 35, 179, 216
- World Health Organization Adverse Reaction Terminology (WHOART), 575–577
- World Health Organization (WHO) antenatal care trial, 178–179

APPENDIX A

TABLES

- Table A.1 Areas of Upper Tail of the Standard Normal Distribution
- Table A.2 Upper Quantiles of a χ^2 Distribution
- Table A.3 Upper Quantiles of a Central t Distribution
- Table A.4 Upper Quantiles of an F Distribution
- Table A.5 Quantiles of the Distribution of Wilcoxon-Mann-Whitney Statistic
- Table A.6 Quantiles of the Wilcoxon Signed Rank Test Statistic
- Table A.7 Tolerance Factor for the Degrees of Confidence $1 - \alpha$
- Table A.8 Upper Quantiles of the Studentized Range Distribution
- Table A.9 Upper Quantiles of the Dennett's Distribution: One-Sided Comparisons with Control

Table A.1 Areas of Upper Tail of the Standard Normal Distribution

<i>z</i>	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139

Source: Table 3 of *Statistical Tables for Science, Engineering and Management*, J. Murdock and J. A. Barnes, Macmillan, London, 1968.

Table A.2 Upper Quantiles of a χ^2 Distribution

$v\alpha$	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	392704.10 ⁻¹⁰	157088.10 ⁻⁹	982069.10 ⁻⁹	393214.10 ⁻⁸	0.0157908	2.70554	3.84146	5.02389	6.63490	7.87944
2	0.0100251	0.0201007	0.0506356	0.102587	0.210720	4.60517	5.99147	7.37776	9.21034	10.5966
3	0.0717212	0.114832	0.297110	0.484419	0.584375	6.25139	7.84743	9.34840	11.3449	12.8381
4	0.206990	0.315795	0.831211	1.145476	1.61031	9.23635	11.0705	12.8705	14.8662	16.7496
5	0.411740	0.554300	1.237347	1.65539	2.20413	10.6446	12.5916	14.4494	16.0863	16.5476
6	0.675727	0.872085	1.239043	1.689887	2.16735	2.83311	12.0170	14.0671	16.0128	18.4753
7	0.989265	1.344419	1.646482	2.17973	2.73264	3.48954	13.3616	15.5073	17.5346	20.2777
8	1.734926	2.087912	2.707039	3.32511	4.16816	4.6837	16.9190	19.0228	21.6660	23.5893
9	2.15585	2.55821	3.24697	3.94030	4.86518	5.9871	15.9871	18.3070	20.4831	23.2093
10	2.60321	3.05347	3.81575	4.57481	5.57779	17.2750	19.6751	21.9200	24.7250	26.7569
11	3.07382	3.57056	4.40379	5.22603	6.30380	18.5494	21.0261	23.3367	26.2170	28.2995
12	3.56503	4.10691	5.00874	5.89186	7.04150	19.8119	22.3621	24.7356	27.6883	29.8194
13	4.07468	4.66043	5.62872	6.57063	7.78953	21.0642	23.6848	26.1190	29.1413	31.3193
14	4.60094	5.22935	6.26214	7.26094	8.54675	22.3072	24.9958	27.5779	30.5779	32.8013
15	5.14224	5.81221	6.90766	7.96164	9.31223	23.5418	26.2962	28.8454	31.9999	34.2672
16	5.69724	6.40776	7.56418	8.67176	10.0852	24.7690	27.5871	30.1910	33.4087	35.7185
17	6.26481	7.01491	8.23075	9.39046	10.8649	25.9894	28.8693	31.5261	34.8053	37.1564
18	6.84398	7.63273	8.90655	10.1170	11.6509	27.2036	30.1435	32.8523	36.1908	38.5822
19	7.43386	8.26040	9.59083	10.8508	12.4426	28.4120	31.4104	34.1696	37.5662	39.9968
20	8.03366	8.89720	10.28293	11.5913	13.2396	29.6151	32.6705	35.4789	38.9321	41.4010
21	8.64272	9.54249	10.9823	12.3380	14.0415	30.8133	33.9244	36.7807	40.2894	42.7956
22	9.26042	10.19567	11.6885	13.0905	14.8479	32.0069	35.1725	38.0757	41.6384	44.1813
23	9.88623	10.8564	12.4011	13.8484	15.6587	33.1963	36.4151	39.3641	42.9798	45.5585
24	10.5197	11.5240	13.1197	14.6114	16.4734	34.3816	37.6525	40.6465	44.3141	46.9278
25	11.1603	12.1981	13.8439	15.3791	17.2919	35.5631	38.8852	41.9232	45.6417	48.2899
26	11.8076	12.8786	14.5733	16.1513	18.1138	36.7412	40.1133	43.1944	46.9630	49.6449
27	12.4613	13.5648	15.3079	16.9279	18.9392	37.9159	41.3372	44.4607	48.2782	50.9933
28	13.1211	14.2565	16.0471	17.7083	19.7677	39.0875	42.5569	45.7222	49.5879	52.3356
29	13.7867	14.9535	16.7908	18.4926	20.5992	40.2560	43.7729	46.9792	50.8922	53.6720
30	20.7065	22.1643	24.4331	26.5093	29.0505	51.8050	55.7585	59.3417	63.6907	66.7659
40	27.9907	29.7067	32.3574	34.7642	37.6886	63.1671	67.5048	71.4202	76.1539	79.4900
50	35.5346	37.4848	40.4817	43.1879	46.4589	74.3970	79.0819	83.2976	88.3794	91.9517
60	43.2752	45.4418	48.7576	51.7393	55.3290	85.5271	90.5312	95.0231	100.4215	104.215
70	51.1720	53.5400	57.1532	60.3915	64.2778	96.5782	101.879	106.629	112.329	116.321
80	59.1963	61.7541	65.6466	69.1260	73.2912	107.5655	113.145	118.136	124.116	128.299
90	67.3276	70.0648	74.2219	77.9295	82.3581	118.498	124.342	129.561	135.807	140.169

Source: Tables of Percentage Points of the χ^2 -Distribution by C. M. Thompson. *Biometrika* (1941). Vol. 32, pp. 188-189.

Table A.3 Upper Quantiles of a Central t Distribution

v/α	0.050	0.025	0.010	0.005
1	6.3138	12.706	25.452	63.647
2	2.9200	4.3027	6.2053	9.9248
3	2.3534	3.1825	4.1765	5.8409
4	2.1318	2.7764	3.4954	4.6041
5	2.0150	2.5706	3.1634	4.0321
6	1.9432	2.4469	2.9687	3.7074
7	1.8946	2.3646	2.8412	3.4995
8	1.8595	2.3060	2.7515	3.3554
9	1.8331	2.2622	2.6850	3.2498
10	1.8125	2.2281	2.6338	3.1693
11	1.7959	2.2010	2.5931	3.1058
12	1.7823	2.1788	2.5600	3.0545
13	1.7709	2.1604	2.5326	3.0123
14	1.7613	2.1448	2.5096	2.9768
15	1.7530	2.1315	2.4899	2.9467
16	1.7459	2.1199	2.4729	2.9208
17	1.7396	2.1098	2.4581	2.8982
18	1.7341	2.1009	2.4450	2.8784
19	1.7291	2.0930	2.4334	2.8609
20	1.7247	2.0860	2.4231	2.8453
21	1.7207	2.0796	2.4138	2.8314
22	1.7171	2.0739	2.4055	2.8188
23	1.7139	2.0687	2.3979	2.8073
24	1.7109	2.0639	2.3910	2.7969
25	1.7081	2.0595	2.3846	2.7874
26	1.7056	2.0555	2.3788	2.7787
27	1.7033	2.0518	2.3734	2.7707
28	1.7011	2.0484	2.3685	2.7633
29	1.6991	2.0452	2.3638	2.7564
30	1.6973	2.0423	2.3596	2.7500
40	1.6839	2.0211	2.3289	2.7045
60	1.6707	2.0003	2.2991	2.6603
120	1.6577	1.9799	2.2699	2.6174
∞	1.6449	1.9600	2.2414	2.5758

Source: Tables of Percentage Points of the t -Distribution by M. Merrington, *Biometrika* (1941), Vol. 32, p. 300.

Table A.4 Upper Quintiles of an F Distribution

		$\alpha = 0.05$																	
v_{2/v_1}	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	249.05	250.09	251.14	252.20	253.25	254.32
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.396	19.413	19.429	19.446	19.454	19.462	19.471	19.479	19.487	19.496	
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8868	8.8452	8.8123	8.7855	8.7446	8.7029	8.6602	8.6385	8.6166	8.5944	8.5720	8.5494	8.5265
4	7.0786	6.9443	6.5914	6.3883	6.2560	6.1631	6.0942	6.0410	5.9988	5.9644	5.9117	5.8578	5.8025	5.7744	5.7459	5.7170	5.6878	5.6581	5.6281
5	6.6079	5.7761	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351	4.6777	4.6188	4.5581	4.5272	4.4957	4.4638	4.4314	4.3984	4.3650
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2066	4.1468	4.0990	4.0600	3.9999	3.9381	3.8742	3.8415	3.8082	3.7743	3.7398	3.7047	3.6688
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5747	3.5108	3.4445	3.4105	3.3758	3.3404	3.3043	3.2674	3.2298
8	5.3177	4.4590	4.0662	3.8378	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2840	3.2184	3.1503	3.1152	3.0794	3.0428	3.0053	2.9669	2.9276
9	5.1174	4.2565	3.8626	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0729	3.0061	2.9365	2.9005	2.8637	2.8259	2.7872	2.7475	2.7067
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130	2.8450	2.7740	2.7372	2.6996	2.6609	2.6211	2.5801	2.5379
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	2.7876	2.7186	2.6464	2.6090	2.5705	2.5309	2.4901	2.4480	2.4045
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	2.6866	2.6169	2.5436	2.5055	2.4663	2.4259	2.3842	2.3410	2.2962
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710	2.6037	2.5331	2.4589	2.4202	2.3803	2.3392	2.2966	2.2524	2.2064
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6021	2.5342	2.4630	2.3879	2.3487	2.3082	2.2664	2.2230	2.1778	2.1307
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437	2.4753	2.4035	2.3275	2.2878	2.2468	2.2043	2.1601	2.1141	2.0658
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377	2.4935	2.4247	2.3522	2.2756	2.2354	2.1938	2.1507	2.1058	2.0589	2.0096
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943	2.4499	2.3807	2.3077	2.2304	2.1898	2.1477	2.1040	2.0584	2.0107	1.9604
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563	2.4117	2.3421	2.2686	2.1906	2.1497	2.1071	2.0629	2.0166	1.9681	1.9168
19	4.3808	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	2.3779	2.3080	2.2341	2.1555	2.1141	2.0712	2.0264	1.9796	1.9302	1.8780
20	4.3513	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928	2.3479	2.2776	2.2033	2.1242	2.0825	2.0391	1.9938	1.9464	1.8963	1.8432
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3661	2.3210	2.2504	2.1757	2.0960	2.0540	2.0102	1.9645	1.9165	1.8657	1.8117
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419	2.2967	2.2258	2.1508	2.0707	2.0283	1.9842	1.9380	1.8895	1.8380	1.7831
23	4.2279	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201	2.2747	2.2036	2.1282	2.0476	2.0050	1.9605	1.9139	1.8649	1.8128	1.7570
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002	2.2547	2.1834	2.1077	2.0236	1.9838	1.9390	1.8920	1.8424	1.7897	1.7331
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	2.2365	2.1649	2.0889	2.0075	1.9643	1.9192	1.8718	1.8217	1.7684	1.7110
26	4.2252	2.3690	2.9751	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655	2.2197	2.1479	2.0716	1.9898	1.9464	1.9010	1.8533	1.8027	1.7488	1.6906
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501	2.2043	2.1323	2.0558	1.9736	1.9299	1.8842	1.8361	1.7851	1.7307	1.6717
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360	2.1900	2.1179	2.0411	1.9586	1.9147	1.8687	1.8203	1.7689	1.7138	1.6541
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2782	2.2229	2.1768	2.1045	2.0275	1.9446	1.9005	1.8543	1.8055	1.7537	1.6881	1.6377
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646	2.0921	2.048	1.9317	1.8874	1.8409	1.7918	1.7396	1.6835	1.6223
40	4.0848	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240	2.0772	2.0032	1.9245	1.8389	1.7929	1.7444	1.6928	1.6373	1.5766	1.5089
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2540	2.1665	2.0970	2.0401	1.9926	1.9174	1.8364	1.7480	1.7001	1.6491	1.5943	1.5343	1.4673	1.3893
120	3.9201	3.0718	2.6802	2.4472	2.2900	2.1750	2.0867	2.0164	1.9588	1.9105	1.8337	1.7505	1.6587	1.6084	1.5543	1.4952	1.4290	1.3519	1.2539
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799	1.8307	1.7522	1.6664	1.5705	1.5173	1.4591	1.3940	1.3180	1.2214	1.0000

Table A.4 (*Continued*)

$\alpha = 0.025$													
$v2/v1$	1	2	3	4	5	6	7	8	9	10	12	15	20
1	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63	976.71	984.87	993.10
2	38.506	39.000	39.165	39.248	39.298	39.331	39.355	39.373	39.387	39.398	39.415	39.448	39.465
3	17.443	16.044	15.439	15.101	14.885	14.735	14.624	14.540	14.473	14.419	14.337	14.253	14.167
4	12.218	10.649	9.9792	9.6045	9.3645	9.1973	9.0741	8.9796	8.9047	8.8439	8.7512	8.6565	8.5599
5	10.007	8.4336	7.7636	7.3879	7.1464	6.9777	6.8531	6.7572	6.6810	6.6192	6.5246	6.4277	6.3285
6	8.8131	7.2598	6.5988	6.2272	5.9876	5.8197	5.6955	5.5996	5.5234	5.4613	5.3662	5.2687	5.1684
7	8.0727	6.5415	5.8898	5.5226	5.2852	5.1186	4.9949	4.8994	4.8232	4.7611	4.6658	4.5678	4.4667
8	7.5709	6.0595	5.4160	5.0526	4.8173	4.6517	4.5286	4.4332	4.3572	4.2951	4.1997	4.1012	3.9995
9	7.2093	5.7147	5.0781	4.7181	4.4844	4.3197	4.1971	4.1020	4.0260	3.9639	3.8682	3.7694	3.6669
10	6.9367	5.4564	4.8256	4.4683	4.2361	4.0721	3.9498	3.8549	3.7790	3.7168	3.6209	3.5217	3.4186
11	6.7241	5.2559	4.6300	4.2751	4.0440	3.8807	3.7586	3.6658	3.5879	3.5257	3.4296	3.3299	3.2261
12	6.5538	5.0959	4.4742	4.1212	3.8911	3.7283	3.6065	3.5118	3.4358	3.3736	3.2773	3.1772	3.0728
13	6.4143	4.9653	4.3472	3.9959	3.7667	3.6043	3.4827	3.3880	3.3120	3.2497	3.1532	3.0527	2.9477
14	6.2979	4.8567	4.2417	3.8919	3.6634	3.5014	3.3799	3.2853	3.2093	3.1469	3.0501	2.9493	2.8437
15	6.1995	4.7650	4.1528	3.8043	3.5764	3.4147	3.2994	3.1987	3.1227	3.0602	2.9633	2.8621	2.7559
16	6.1151	4.6867	4.0768	3.7294	3.5021	3.3406	3.2194	3.1248	3.0488	2.9862	2.8890	2.7875	2.6808
17	6.0420	4.6189	4.0112	3.6648	3.4379	3.2767	3.1556	3.0610	2.9849	2.9222	2.8249	2.7230	2.6158
18	5.9781	4.5597	3.9539	3.7505	3.5346	3.3220	3.0999	3.0053	2.9291	2.8664	2.7689	2.6667	2.5590
19	5.9216	4.5075	3.9034	3.5587	3.3327	3.1718	3.0509	2.9563	2.8800	2.8173	2.7196	2.6171	2.5089
20	5.8715	4.4613	3.8887	3.5147	3.2891	3.1283	3.0074	2.9128	2.8365	2.7737	2.6758	2.5731	2.4645
21	5.8266	4.4199	3.8188	3.4754	3.2501	3.0895	2.9686	2.8740	2.7977	2.7348	2.6368	2.5338	2.4547
22	5.7863	4.3828	3.7829	3.4401	3.2151	3.0546	2.9338	2.8392	2.7628	2.6998	2.6017	2.4984	2.3890
23	5.7498	4.3492	3.7505	3.4083	3.1835	3.0232	2.9024	2.8077	2.7313	2.6682	2.5699	2.4665	2.3567
24	5.7167	4.3187	3.7211	3.3794	3.1548	2.9946	2.8738	2.7791	2.7027	2.6396	2.5412	2.4374	2.3273
25	5.6864	4.2909	3.6943	3.3530	3.1287	2.9685	2.8478	2.7531	2.6766	2.6135	2.5149	2.4110	2.3005
26	5.6586	4.2655	3.6697	3.3289	3.1048	2.9447	2.8240	2.7293	2.6528	2.5895	2.4909	2.3867	2.2759
27	5.6331	4.2421	3.6472	3.3067	3.0828	2.9228	2.8021	2.7074	2.6309	2.5676	2.4688	2.3644	2.2533
28	5.6096	4.2205	3.6264	3.2863	3.0625	2.9027	2.7820	2.6872	2.6106	2.5473	2.4484	2.3438	2.2324
29	5.5878	4.2006	3.6072	3.2674	3.0438	2.8840	2.7633	2.6686	2.5919	2.5286	2.4295	2.3248	2.2131
30	5.5675	4.1821	3.5894	3.2499	3.0265	2.8667	2.7460	2.6513	2.5746	2.5112	2.4210	2.3072	2.1952
40	5.4239	4.0510	3.4633	3.1261	2.9037	2.7444	2.6238	2.5289	2.4519	2.3882	2.2882	2.1819	2.0677
60	5.2857	3.9253	3.3425	3.0077	2.7863	2.5068	2.4117	2.3344	2.2702	2.1692	2.0613	1.9445	1.8152
120	5.1524	3.8046	3.2270	2.8943	2.6740	2.5154	2.3948	2.2994	2.2217	2.1570	2.0548	1.9450	1.8249
∞	5.0239	3.6889	3.1161	2.7858	2.5665	2.4082	2.2875	2.1918	2.1136	2.0483	1.9447	1.8326	1.7085

$\alpha = 0.010$																			
v2/v1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052.2	4999.5	5403.3	5624.6	5763.7	5839.0	5928.3	5981.6	6022.5	6055.8	6106.3	6157.3	6208.7	6234.6	6260.7	6286.8	6313.0	6339.4	6366.0
2	98.503	99.000	99.166	99.249	99.322	99.356	99.374	99.388	99.399	99.416	99.432	99.449	99.458	99.466	99.474	99.483	99.491	99.501	
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229	27.052	26.872	26.690	26.598	26.505	26.411	26.316	26.221	26.125
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.374	14.198	14.020	13.929	13.838	13.745	13.652	13.558	13.463
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.8883	9.7222	9.5527	9.4665	9.3793	9.2912	9.2020	9.1118	9.0204
6	13.745	10.925	9.7795	9.1483	8.7459	8.4661	8.2600	8.1016	7.9761	7.8741	7.7183	7.5590	7.3958	7.3127	7.2285	7.1432	7.0568	6.9690	6.8801
7	12.246	9.5466	8.4513	7.8467	7.4604	7.1914	6.9928	6.8401	6.7188	6.6201	6.6491	6.3143	6.1554	6.0743	5.9921	5.9084	5.8236	5.7372	5.6495
8	11.259	8.6491	7.5910	6.7060	6.6318	6.3707	6.1776	6.0289	5.9106	5.8143	5.6668	5.5151	5.3591	5.2793	5.1981	5.1156	5.0316	4.9460	4.8588
9	10.561	8.0215	6.9919	6.4221	6.0569	5.8018	5.6129	5.4671	5.3511	5.2565	5.1114	4.9621	4.8080	4.7290	4.6486	4.5667	4.4831	4.3978	4.3105
10	10.044	7.5594	6.5523	5.9943	5.6363	5.3858	5.2001	5.0567	4.9424	4.8492	4.7059	4.5582	4.4054	4.3269	4.2469	4.1653	4.0819	3.9965	3.9090
11	9.6460	7.2057	6.2167	5.6683	5.3160	5.0692	4.8861	4.7445	4.6315	4.5393	4.3974	4.2509	4.0990	4.0209	3.9411	3.8596	3.7761	3.6904	3.6025
12	9.302	6.9266	5.9226	5.4119	5.0643	4.8206	4.6395	4.4994	4.3875	4.2961	4.1553	4.0096	3.8584	3.7805	3.7008	3.6192	3.5355	3.4494	3.3608
13	9.0738	6.7010	5.7394	5.2053	4.8616	4.6204	4.4410	4.3021	4.1911	4.1003	3.9603	3.8154	3.6646	3.5868	3.5070	3.4253	3.3413	3.2548	3.1654
14	8.8616	6.5149	5.5639	5.0354	4.6950	4.4558	4.2779	4.1399	4.0297	3.9394	3.8001	3.6557	3.5052	3.4274	3.3476	3.2556	3.1813	3.0942	3.0040
15	8.6831	6.3589	5.4170	4.8932	4.5556	4.3183	4.1415	4.0045	3.8948	3.8049	3.6662	3.5222	3.3719	3.2940	3.2141	3.1319	3.0471	2.9595	2.8684
16	8.5310	6.2262	5.2922	4.7726	4.4374	4.2016	4.0259	3.8896	3.7804	3.6909	3.5527	3.4089	3.2588	3.1808	3.1007	3.0182	2.9330	2.8447	2.7528
17	8.3997	6.1121	5.1850	4.6690	4.3359	4.1015	3.9267	3.7910	3.6822	3.5931	3.4552	3.3117	3.1615	3.0835	3.0032	2.9205	2.8348	2.7459	2.6530
18	8.2854	6.0129	5.0919	4.5790	4.2479	4.0146	3.8406	3.7054	3.5971	3.5082	3.3706	3.2273	3.0771	2.990	2.9185	2.8354	2.7493	2.6597	2.5660
19	8.1850	5.9259	5.0103	4.5003	4.1708	3.9386	3.7653	3.6305	3.5225	3.4338	3.2965	3.1533	3.0031	2.9249	2.8442	2.7608	2.6742	2.5839	2.4893
20	8.0960	5.8489	4.9382	4.4307	4.1027	3.8714	3.6987	3.5644	3.4567	3.3682	3.2311	3.0880	2.9377	2.8594	2.7785	2.6947	2.6077	2.5168	2.4212
21	8.0166	5.7804	4.8740	4.3688	4.0421	3.8117	3.6396	3.5056	3.3981	3.3098	3.1729	3.0299	2.8796	2.8011	2.7200	2.6359	2.5484	2.4568	2.3603
22	7.9454	5.7190	4.8166	4.3134	3.9880	3.7583	3.5867	3.4530	3.3458	3.2576	3.1209	2.9780	2.8274	2.7488	2.6675	2.5831	2.4951	2.4029	2.3055
23	7.8811	5.6637	4.7649	4.2635	3.9392	3.7102	3.5390	3.4057	3.2986	3.2106	3.0740	2.9311	2.7805	2.7017	2.6202	2.5355	2.4471	2.3542	2.2559
24	7.8229	5.6136	4.7181	4.2184	3.8951	3.6667	3.4959	3.3629	3.2560	3.1681	3.0316	2.8887	2.7380	2.6591	2.5773	2.4923	2.4035	2.3099	2.2107
25	7.7698	5.5680	4.6755	4.1774	3.8550	3.6272	3.4568	3.3239	3.2172	3.1294	3.0931	2.8502	2.6993	2.6203	2.5383	2.4530	2.3637	2.2695	2.1694
26	7.7213	5.5263	4.6366	4.1400	3.8183	3.5911	3.4210	3.2884	3.1818	3.0941	2.9579	2.8150	2.6640	2.5848	2.5026	2.4170	2.3273	2.2325	2.1315
27	7.6767	5.4881	4.6009	4.1056	3.7848	3.5580	3.3882	3.2558	3.1494	3.0618	2.9256	2.7827	2.6316	2.5522	2.4699	2.3840	2.2938	2.1984	2.0965
28	7.6356	5.4529	4.5681	4.0740	3.7539	3.5276	3.3581	3.2259	3.1195	3.0320	2.8959	2.7530	2.6017	2.5223	2.4397	2.3535	2.2629	2.1670	2.0642
29	7.5976	5.4205	4.5378	4.0449	3.7254	3.4995	3.3302	3.1982	3.0920	3.0045	2.8685	2.7256	2.5742	2.4946	2.4118	2.3253	2.2344	2.1378	2.0342
30	7.5625	5.3904	4.5097	4.0179	3.6990	3.4735	3.3045	3.1726	3.0665	2.9791	2.8431	2.7002	2.5487	2.4689	2.3860	2.2992	2.2079	2.1107	2.0062
31	7.3141	5.1785	4.3126	3.8283	3.5138	3.2910	3.1238	2.9930	2.8876	2.8005	2.6648	2.5216	2.3689	2.2880	2.2034	2.1142	2.0194	1.9172	1.8047
32	7.0771	4.9774	4.1259	3.6491	3.3389	3.1187	2.9530	2.8233	2.7185	2.6318	2.4961	2.3523	2.1978	2.1154	2.0285	1.9360	1.8363	1.7263	1.6006
33	6.8510	4.7865	3.9493	3.4796	3.1735	2.9559	2.7918	2.6629	2.5586	2.4721	2.3363	2.1915	2.0346	1.9500	1.8600	1.7628	1.6557	1.5330	1.3805
34	6.6349	4.6052	3.7816	3.3192	3.0173	2.8020	2.6393	2.5113	2.4073	2.3209	2.1848	2.0385	1.8783	1.7908	1.6964	1.5923	1.4730	1.3246	1.0000

Table A.4 (Continued)

$\alpha = 0.005$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	16211	20000	21615	22500	23056	23437	23715	23925	24091	24224	24426	24630	24836	24940	25044	25148	25253	25359	25465
2	198.50	199.00	199.17	199.25	199.30	199.33	199.36	199.39	199.40	199.42	199.43	199.45	199.46	199.47	199.48	199.49	199.49	199.51	
3	55.552	49.799	47.467	46.195	45.392	44.838	41.434	44.126	43.882	43.686	43.387	43.085	42.778	42.622	42.466	42.308	42.149	41.989	41.829
4	31.333	26.284	24.259	23.155	22.456	21.975	21.352	21.139	20.967	20.705	20.438	20.167	20.030	19.892	19.752	19.611	19.468	19.325	
5	22.785	18.314	16.530	15.556	14.940	14.513	14.200	13.961	13.722	13.618	13.384	13.146	12.903	12.780	12.656	12.530	12.402	12.274	12.161
6	18.635	14.544	12.917	12.028	11.464	11.073	10.786	10.566	10.391	10.250	10.034	9.8140	9.5888	9.4741	9.3583	9.2408	9.1219	9.0015	8.8793
7	16.236	12.404	10.882	10.050	9.5221	9.1554	8.8854	8.6781	8.5138	8.3803	8.1764	7.9678	7.7540	7.6450	7.5345	7.4225	7.3088	7.1933	7.0760
8	14.688	11.042	9.5965	8.051	8.3018	7.9320	7.6942	7.4960	7.3386	7.2107	7.0149	6.8143	6.6082	6.5029	6.3961	6.2875	6.1772	6.0648	5.9505
9	13.614	10.107	8.7171	7.9559	7.4711	7.1338	6.8849	6.6933	6.5411	6.4171	6.2274	6.0325	5.8318	5.7292	5.6248	5.5186	5.4104	5.3001	5.1875
10	12.826	9.4270	8.0807	7.3428	6.8723	6.5446	6.3025	6.1159	5.9676	5.8467	5.6613	5.4707	5.2740	5.1732	5.0705	4.9659	4.8592	4.7501	4.6385
11	12.226	8.9122	7.6004	6.8809	6.4217	6.1015	5.8648	5.6821	5.5368	5.4182	5.2363	5.0489	4.8552	4.7557	4.6543	4.5508	4.4450	4.3367	4.2256
12	11.754	8.5096	7.2258	6.5215	6.0711	5.7570	5.5245	5.3451	5.2021	5.0855	4.9063	4.7214	4.5299	4.4315	4.3309	4.2282	4.1229	4.0149	3.9039
13	11.374	8.1805	6.9257	6.2335	5.7910	5.4819	5.2529	5.0761	4.9551	4.8199	4.6429	4.4600	4.2703	4.1726	4.0727	3.9704	3.8655	3.7577	3.6465
14	11.060	7.9217	6.6803	5.9984	5.5623	5.2574	5.0313	4.8566	4.7173	4.6034	4.4281	4.2468	4.0585	3.9614	3.8619	3.7600	3.6553	3.5473	3.4359
15	10.798	7.7008	6.4760	5.8029	5.3721	5.0708	4.8473	4.6743	4.5664	4.4236	4.2498	4.0698	3.8826	3.7559	3.6867	3.5850	3.4803	3.3722	3.2602
16	10.575	7.5138	6.3034	5.6378	5.2117	4.9134	4.6920	4.5207	4.3838	4.2719	4.0994	3.9205	3.7342	3.6378	3.5388	3.4372	3.3324	3.2240	3.1115
17	10.384	7.3536	6.1556	5.4967	5.0746	4.7789	4.5594	4.3893	4.2535	4.1423	3.9709	3.7929	3.6073	3.5112	3.4124	3.3107	3.2058	3.0971	2.9839
18	10.218	7.2148	6.0277	5.3746	4.9560	4.6627	4.4448	4.2759	4.1410	4.0305	3.8599	3.6827	3.4977	3.4017	3.3030	3.2014	3.0962	2.9871	2.8732
19	10.073	7.0935	5.9161	5.2681	4.8526	4.5614	4.3448	4.1770	4.0428	3.9329	3.7631	3.5866	3.4020	3.3062	3.2075	3.1058	3.0004	2.8908	2.7762
20	19.9439	6.9865	5.8177	5.1743	4.7616	4.4721	4.2569	4.0900	3.9564	3.8470	3.6779	3.5020	3.3178	3.2220	3.1234	3.0215	2.9159	2.8058	2.6904
21	19.8295	6.8914	5.7304	5.0911	4.6808	4.3931	4.1789	4.0128	3.8799	3.7709	3.6024	3.4270	3.2431	3.1474	3.0488	2.9467	2.8408	2.7302	2.6140
22	9.7271	6.8804	5.6524	5.0168	4.6088	4.3225	4.1094	3.9440	3.8116	3.7030	3.5350	3.3600	3.1764	3.0807	2.9821	2.8776	2.76625	2.5455	
23	9.6348	6.7300	5.5823	4.9500	4.5441	4.2459	3.8822	3.7502	3.6420	3.4745	3.2999	3.1165	3.0208	2.9221	2.8198	2.7132	2.6016	2.4837	
24	9.5513	6.6610	5.5190	4.8898	4.4857	4.2019	3.9905	3.8264	3.6949	3.5870	3.4117	3.2456	3.0624	2.9067	2.8679	2.7654	2.6385	2.5463	2.4276
25	9.4753	6.5982	5.4615	4.8351	4.4327	4.1500	3.9394	3.7758	3.6447	3.5370	3.3704	3.1963	3.0133	2.9176	2.8187	2.7160	2.6088	2.4960	2.3765
26	9.4059	6.5409	4.7852	4.3844	4.1027	3.8928	3.7297	3.5989	3.4916	3.3252	3.1515	2.9685	2.8728	2.7738	2.6709	2.5633	2.4501	2.3297	
27	9.3423	6.4885	5.3611	4.7396	4.3402	4.0594	3.8501	3.6875	3.5571	3.4499	3.2839	3.1104	2.9275	2.8318	2.7327	2.6296	2.5217	2.4078	2.2867
28	9.2838	6.4403	5.3170	4.6977	4.2996	4.0197	3.8110	3.6487	3.5186	3.4117	3.2460	3.0727	2.8899	2.7941	2.6949	2.5916	2.4834	2.3689	2.2469
29	9.2297	6.3958	5.2764	4.6591	4.2622	3.9830	3.7749	3.6130	3.4832	3.3765	3.2111	3.0379	2.8551	2.7594	2.6601	2.5565	2.4479	2.3330	2.2102
30	9.1797	6.3547	5.2388	4.6233	4.2276	3.9492	3.7416	3.5801	3.4505	3.3440	3.1787	3.0057	2.8230	2.7272	2.6278	2.5241	2.4151	2.2997	2.1760
40	8.8278	6.0664	4.9759	4.3738	3.9860	3.7129	3.5088	3.3498	3.2220	3.1167	2.9531	2.7811	2.5984	2.5020	2.4015	2.2958	2.1838	2.0635	1.9318
60	8.4946	5.7950	4.7290	4.1399	3.7600	3.2911	3.1344	3.0083	2.9042	2.7419	2.5705	2.3872	2.2898	2.1874	2.0789	1.9622	1.8341	1.6885	
120	8.1790	5.5393	4.4973	3.9207	3.5482	3.2849	3.0874	2.9330	2.8083	2.7052	2.5439	2.3727	2.1881	2.0890	1.9839	1.8709	1.7469	1.6055	1.4311
∞	7.8794	5.2983	4.2794	3.7151	3.3499	3.0913	2.8968	2.7444	2.6210	2.5188	2.3583	2.1868	1.9998	1.8983	1.7891	1.6691	1.5325	1.3637	1.0000

Source: Tables of Percentage Points of the Inverted beta (F)-Distribution by M. Merrington and C. M. Thompson, *Biometrika* (1942), Vol. 33, pp. 73–88.

Table A.5 Quantiles of the Distribution of Wilcoxon–Mann–Whitney Statistic

n_1	α	$n_2 = 2$	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	0.001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
	0.01	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	2	2	2
	0.025	0	0	0	0	0	0	0	1	1	1	2	2	2	2	2	3	3	3	3
	0.05	0	0	0	1	1	1	2	2	2	2	3	3	4	4	4	4	5	5	5
	0.10	0	1	1	2	2	2	3	3	4	4	5	5	5	6	6	7	7	8	8
3	0.001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
	0.005	0	0	0	0	0	0	0	0	1	1	2	2	2	2	3	3	3	4	4
	0.01	0	0	0	0	0	0	1	1	2	2	2	3	3	4	4	4	5	5	6
	0.025	0	0	0	0	1	2	2	3	3	4	4	5	5	6	6	7	7	8	9
	0.05	0	1	1	2	3	3	4	5	5	6	6	7	8	8	9	10	10	11	12
	0.10	1	2	2	3	4	5	6	6	7	8	9	10	11	11	12	13	14	15	16
4	0.001	0	0	0	0	0	0	0	0	1	1	1	2	2	2	3	3	4	4	4
	0.005	0	0	0	0	0	1	1	2	2	3	4	4	5	5	6	6	7	7	9
	0.01	0	0	0	1	2	3	4	5	5	6	7	8	9	10	11	12	12	13	15
	0.025	0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	15	16	17	18
	0.05	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	19
	0.010	1	2	4	5	6	7	8	10	11	12	13	14	16	17	18	19	21	22	23
5	0.001	0	0	0	0	0	0	1	2	2	3	4	5	6	7	8	8	9	10	11
	0.005	0	0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	14
	0.01	0	0	1	2	3	4	5	6	7	8	9	10	12	13	14	15	16	17	17
	0.025	0	1	2	3	4	6	7	8	9	10	12	13	14	15	16	18	19	20	21
	0.05	1	2	3	5	6	7	9	10	12	13	14	16	17	19	20	21	23	24	26
	0.10	2	3	5	6	8	9	11	13	14	16	18	19	21	23	24	26	28	29	31

Table A.5 (Continued)

n_1	α	$n_2 = 2$	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
6	0.001	0	0	0	0	0	2	3	4	5	5	6	7	8	9	10	11	12	13	13
	0.005	0	0	1	2	3	4	5	6	7	8	10	11	12	13	14	16	17	18	19
	0.01	0	0	2	3	4	5	6	7	8	9	10	12	13	14	16	17	19	20	23
	0.025	0	2	3	4	6	7	9	11	12	14	15	17	18	20	22	23	25	26	28
	0.05	1	3	4	6	8	9	11	13	15	17	18	20	22	24	26	27	29	31	33
	0.10	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	35	37	39
7	0.001	0	0	0	0	1	2	3	4	6	7	8	9	10	11	12	14	15	16	17
	0.005	0	0	1	2	4	5	7	8	10	11	13	14	16	17	19	20	22	23	25
	0.01	0	1	2	4	5	7	8	10	12	13	15	17	18	20	22	24	25	27	29
	0.025	0	2	4	6	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35
	0.05	1	3	5	7	9	12	14	16	18	20	22	25	27	29	31	34	36	38	40
	0.10	2	5	7	9	12	14	17	19	22	24	27	29	32	34	37	39	42	44	47
8	0.001	0	0	0	1	2	3	5	6	7	9	10	12	14	16	18	19	21	22	22
	0.005	0	0	1	2	3	5	7	8	10	12	14	16	18	21	23	25	27	29	31
	0.01	0	1	3	5	7	9	11	14	16	18	20	23	25	27	30	32	35	37	39
	0.025	1	3	5	7	9	11	14	16	19	21	24	27	29	32	34	37	40	42	42
	0.05	2	4	6	9	11	14	16	19	21	24	27	29	31	34	37	40	42	45	48
	0.10	3	6	8	11	14	17	20	23	25	28	31	34	37	40	43	46	49	52	55
9	0.001	0	0	0	2	3	4	6	8	9	11	13	15	16	18	20	22	24	26	27
	0.005	0	1	2	4	6	8	10	12	14	17	19	22	24	27	29	32	34	37	39
	0.01	0	2	4	6	8	10	12	15	17	19	22	24	27	29	32	35	38	40	43
	0.025	1	3	5	8	11	13	16	18	21	24	27	29	31	34	37	40	43	46	49
	0.05	2	5	7	10	13	16	19	22	25	28	31	34	37	40	43	46	49	52	55
	0.10	3	6	10	13	16	19	23	26	29	32	36	39	42	46	49	53	56	59	63
10	0.001	0	0	1	2	4	6	7	9	11	13	15	18	20	22	24	26	28	30	33
	0.005	0	1	3	5	7	10	12	14	17	20	23	25	28	31	34	37	39	42	43
	0.01	0	2	4	7	9	12	14	17	20	23	25	28	31	34	37	40	43	45	48
	0.025	1	4	6	9	12	15	18	21	24	27	30	34	37	40	43	46	49	53	56
	0.05	2	5	8	12	15	18	21	25	28	32	35	38	42	45	49	52	56	59	63
	0.10	4	7	11	14	18	22	25	29	33	37	40	44	48	52	55	59	63	67	71

0.001	0	0	1	3	5	7	9	11	13	16	18	21	23	25	28	30	33	38	
0.005	0	0	1	2	5	8	10	13	16	19	22	25	28	32	35	38	40	43	49
0.01	0	0	1	4	7	10	14	17	20	24	27	31	34	38	41	45	48	51	54
0.025	1	0	4	7	10	14	17	20	24	28	32	35	39	43	47	51	56	59	63
0.05	0	2	6	9	13	17	20	24	28	32	35	39	43	47	51	55	58	62	66
0.10	4	8	12	16	20	24	28	32	37	41	45	49	53	58	62	66	70	74	79
11																			
12																			
0.001	0	0	1	3	5	8	10	13	15	18	21	24	26	29	32	35	38	41	43
0.005	0	0	2	4	7	10	13	16	19	22	25	28	32	36	39	43	47	50	52
0.01	0	0	3	6	9	12	15	18	22	25	29	32	36	39	43	47	50	54	57
0.025	2	5	8	12	15	19	23	27	30	34	38	42	46	50	54	58	62	66	70
0.05	3	6	10	14	18	22	27	31	35	39	43	48	52	56	61	65	69	73	78
0.10	5	9	13	18	22	27	31	36	40	45	50	54	59	64	68	73	78	82	87
13																			
14																			
0.001	0	0	0	2	4	6	9	12	15	18	21	24	27	30	33	36	39	43	49
0.005	0	0	2	4	8	11	14	18	21	25	28	32	35	39	43	46	50	54	61
0.01	1	3	6	10	13	17	21	24	28	32	36	40	44	48	52	56	60	64	68
0.025	2	5	9	13	17	21	25	29	34	38	42	46	51	55	60	64	68	73	77
0.05	3	7	11	16	20	25	29	34	38	43	48	52	57	62	66	71	76	81	85
0.10	5	10	14	19	24	29	34	39	44	49	54	59	64	69	75	80	85	90	95
15																			

Table A.5 (Continued)

n_1	α	$n_2 = 2$	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
16	0.001	0	0	3	6	9	12	16	20	24	28	32	36	40	44	49	53	57	61	66
	0.005	0	3	6	10	14	19	23	28	32	37	42	46	51	56	61	66	71	75	80
	0.01	1	4	8	13	17	22	27	32	37	42	47	52	57	62	67	72	77	83	88
	0.025	2	7	12	16	22	27	32	38	43	48	54	60	65	71	76	82	87	93	99
	0.05	4	9	15	20	26	31	37	43	49	55	61	66	72	78	84	90	96	102	108
	0.10	6	12	18	24	30	37	43	49	55	62	68	75	81	87	94	100	107	113	120
17	0.001	0	1	3	6	10	14	18	22	26	30	35	39	44	48	53	58	62	67	71
	0.005	0	3	7	11	16	20	25	30	35	40	45	50	55	61	66	71	76	82	87
	0.01	1	5	9	14	19	24	29	34	39	45	50	56	61	67	72	78	83	89	94
	0.025	3	7	12	18	23	29	35	40	46	52	58	64	70	76	82	88	94	100	106
	0.05	4	10	16	21	27	34	40	46	52	58	65	71	78	84	90	97	103	110	116
	0.10	7	13	19	26	32	39	46	53	59	66	73	80	86	93	100	107	114	121	128
18	0.001	0	1	4	7	11	15	19	24	28	33	38	43	47	52	57	62	67	72	77
	0.005	0	3	7	12	17	22	27	32	38	43	48	54	59	65	71	76	82	88	93
	0.01	1	5	10	15	20	25	31	37	42	48	54	60	66	71	77	83	89	95	101
	0.025	3	8	13	19	25	31	37	43	49	56	62	68	75	81	87	94	100	107	113
	0.05	5	10	17	23	29	36	42	49	56	62	69	76	83	89	96	103	110	117	124
	0.10	7	14	21	28	35	42	49	56	63	70	78	85	92	99	107	114	121	129	136
19	0.001	0	1	4	8	12	16	21	26	30	35	41	46	51	56	61	67	72	78	83
	0.005	1	4	8	13	18	23	29	34	40	46	52	58	64	70	75	82	88	94	100
	0.01	2	5	10	16	21	27	33	39	45	51	57	64	70	76	83	89	95	102	108
	0.025	3	8	14	20	26	33	39	46	53	59	66	73	79	86	93	100	107	114	120
	0.05	5	11	18	24	31	38	45	52	59	66	73	81	88	95	102	110	117	124	131
	0.10	8	15	22	29	37	44	52	59	67	74	82	90	98	105	113	121	129	136	144
20	0.001	0	1	4	8	13	17	22	27	33	38	43	49	55	60	66	71	77	83	89
	0.005	1	4	9	14	19	25	31	37	43	49	55	61	68	74	80	87	93	100	106
	0.01	2	6	11	17	23	29	35	41	48	54	61	68	74	81	88	94	101	108	115
	0.025	3	9	15	21	28	35	42	49	56	63	70	77	84	91	99	106	113	120	128
	0.05	5	12	19	26	33	40	48	55	63	70	78	85	93	101	108	116	124	131	139
	0.10	8	16	23	31	39	47	55	63	71	79	87	95	103	111	120	128	136	144	152

Source: Table 1 of Extended Tables of Critical Values for Wilcoxon's Test Statistic by L. R. Verdooren, *Biometrika* (1963), Vol. 50, pp. 177-186.

Table A.6 Quantiles of the Wilcoxon Signed-Rank Test Statistic^a

<i>n</i>	<i>W</i> _{0.005}	<i>W</i> _{0.01}	<i>W</i> _{0.025}	<i>W</i> _{0.05}	<i>W</i> _{0.10}	<i>W</i> _{0.20}	<i>W</i> _{0.30}	<i>W</i> _{0.40}	<i>W</i> _{0.50}	$\frac{n(n+1)}{2}$
4	0	0	0	0	1	3	3	4	5	10
5	0	0	0	1	3	4	5	6	7.5	15
6	0	0	1	3	4	6	8	9	10.5	21
7	0	1	3	4	6	9	11	12	14	28
8	1	2	4	6	9	12	14	16	18	36
9	2	4	6	9	11	15	18	20	22.5	45
10	4	6	9	11	15	19	22	25	27.5	55
11	6	8	11	14	18	23	27	30	33	66
12	8	10	14	18	22	28	32	36	39	78
13	10	13	18	22	27	33	38	42	45.5	91
14	13	16	22	26	32	39	44	48	52.5	105
15	16	20	26	31	37	45	51	55	60	120
16	20	24	30	36	43	51	58	63	68	136
17	24	28	35	42	49	58	65	71	76.5	153
18	28	33	41	48	56	66	73	80	85.5	171
19	33	38	47	54	63	74	82	89	95	190
20	38	44	53	61	70	83	91	98	105	210
21	44	50	59	88	78	91	100	108	115.5	131
22	49	56	67	76	87	100	110	119	126.5	153
23	55	63	74	84	95	110	120	130	138	176
24	62	70	82	92	105	120	131	141	150	300
25	69	77	90	101	114	131	143	153	162.5	325
26	76	85	99	111	125	142	155	165	175.5	351
27	84	94	108	120	135	154	167	178	189	378
28	92	102	117	131	146	166	180	192	203	406
29	101	111	127	141	158	178	193	206	217.5	435
30	110	121	138	152	170	191	207	220	232.5	465
31	119	131	148	164	182	205	221	235	248	496
32	129	141	160	176	195	219	236	250	264	528
33	139	152	171	188	208	233	251	266	280.5	561
34	149	163	183	201	222	248	266	282	297.5	595
35	160	175	196	214	236	263	283	299	315	630
36	172	187	209	228	251	279	299	317	333	666
37	184	199	222	242	266	295	316	335	351.5	703
38	196	212	236	257	282	312	334	353	370.5	741
39	208	225	250	272	298	329	352	372	390	780
40	221	239	265	287	314	347	371	391	410	820
41	235	253	280	303	331	365	390	411	430.5	861
42	248	267	295	320	349	384	409	431	451.5	903
43	263	282	311	337	366	403	429	452	473	946
44	277	297	328	354	385	422	450	473	495	990
45	292	313	344	372	403	442	471	495	517.5	1035
46	308	329	362	390	423	463	492	517	540.5	1081
47	324	346	379	408	442	484	514	540	564	1128
48	340	363	397	428	463	505	536	563	588	1176
49	357	381	416	447	483	527	559	587	612.5	1225
50	374	398	435	467	504	550	583	611	637.5	1275

Source: Adapted from Harder and Owen (1970), with permission from the Institute of Mathematical Statistics.

^aFor *n* larger than 50, the *p*th quartile ω_p of the Wilcoxon signed-rank test statistic may be approximated by $\omega_p = [n(n+1)/4] + \chi_p \sqrt{n(n+1)(2n+1)/24}$, where χ_p is the *p*th quartile of a standard normal random variable, obtained from Appendix A.1.

Table A.7 Tolerance Factor For the Degrees Of Confidence $1 - \alpha$

n	$1 - \alpha = 0.95$			$1 - \alpha = 0.99$		
	0.90	0.95	0.99	0.90	0.95	0.99
2	32.019	37.674	48.430	160.193	188.491	242.300
3	8.380	9.916	12.861	18.930	22.401	29.055
4	5.369	6.370	8.299	9.398	11.150	14.527
5	4.275	5.079	6.634	6.612	7.855	10.260
6	3.712	4.414	5.775	5.337	6.345	8.301
7	3.369	4.007	5.248	4.613	5.488	7.187
8	3.136	3.732	4.891	4.147	4.936	6.468
9	2.967	3.532	4.631	3.822	4.550	5.966
10	2.839	3.379	4.433	3.582	4.265	5.594
11	2.737	3.259	4.277	3.397	4.045	5.308
12	2.655	3.162	4.150	3.250	3.870	5.079
13	2.587	3.081	4.044	3.130	3.727	4.893
14	2.529	3.012	3.955	3.029	3.608	4.737
15	2.480	2.954	3.878	2.945	3.507	4.605
16	2.437	2.903	3.812	2.872	3.421	4.492
17	2.400	2.858	3.754	2.808	3.345	4.393
18	2.366	2.819	3.702	2.753	3.279	4.307
19	2.337	2.784	3.656	2.703	3.221	4.230
20	2.310	2.752	3.615	2.659	3.168	4.161
25	2.208	2.631	3.457	2.494	2.972	3.904
30	2.140	2.549	3.350	2.385	2.841	3.733
35	2.090	2.490	3.272	2.306	2.748	3.611
40	2.052	2.445	3.213	2.247	2.677	3.518
45	2.021	2.408	3.165	2.200	2.621	3.444
50	1.996	2.379	3.126	2.162	2.576	3.385
55	1.976	2.354	3.094	2.130	2.538	3.335
60	1.958	2.333	3.066	2.103	2.506	3.293
65	1.943	2.315	3.042	2.080	2.478	3.257
70	1.929	2.299	3.021	2.060	2.454	3.225
75	1.917	2.285	3.002	2.042	2.433	3.197
80	1.907	2.272	2.986	2.026	2.414	3.173
85	1.897	2.261	2.971	2.012	2.397	3.150
90	1.889	2.251	2.958	1.999	2.382	3.130
95	1.881	2.241	2.945	1.967	2.368	3.112
100	1.874	2.233	2.934	1.977	2.355	3.096
150	1.825	2.175	2.859	1.905	2.270	2.983
200	1.798	2.143	2.816	1.865	2.222	2.921
250	1.780	2.121	2.788	1.839	2.191	2.880
300	1.767	2.106	2.767	1.820	2.169	2.850
400	1.749	2.084	2.739	1.794	2.138	2.809
500	1.737	2.070	2.721	1.777	2.117	2.783
600	1.729	2.060	2.707	1.764	2.102	2.763
700	1.722	2.052	2.697	1.755	2.091	2.748
800	1.717	2.046	2.688	1.747	2.082	2.736
900	1.712	2.040	2.682	1.741	2.075	2.726
1000	1.709	2.036	2.676	1.736	2.068	2.718
∞	1.645	1.960	2.576	1.645	1.960	2.576

Source: Adapted by permission from *Techniques of Statistical Analysis* by C. Eisenhart, M. W. Hastay, and W. A. Wallis. Copyright 1947, McGraw-Hill Book Company, Inc.

Table A.8 Upper Quantiles of the Studentized Range Distribution

v	α	<i>t</i>									
		2	3	4	5	6	7	8	9	10	
(a) $t = 2\text{--}10$											
1	.20	4.353	6.615	8.075	9.138	9.966	10.64	11.21	11.70	12.12	
	.10	8.929	13.44	16.36	18.49	20.15	21.51	22.64	23.62	24.48	
	.05	17.97	26.98	32.82	37.08	40.41	43.12	45.4	47.36	49.07	
	.01	90.03	135.0	164.3	185.6	202.2	215.8	227.2	237.0	245.6	
2	.20	2.667	3.820	4.559	5.098	5.521	5.867	6.158	6.409	6.630	
	.10	4.130	5.733	6.773	7.538	8.139	8.633	9.049	9.409	9.725	
	.05	6.085	8.331	9.798	10.88	11.74	12.44	13.03	13.54	13.99	
	.01	14.04	19.02	22.29	24.72	26.63	28.20	29.53	30.68	31.69	
3	.20	2.316	3.245	3.833	4.261	4.597	4.872	5.104	5.305	5.481	
	.10	3.328	4.467	5.199	5.738	6.162	6.511	6.806	7.062	7.287	
	.05	4.501	5.910	6.825	7.502	8.037	8.478	8.853	9.177	9.462	
	.01	8.261	10.62	12.17	13.33	14.24	15.00	15.64	16.20	16.69	
4	.20	2.168	3.004	3.527	3.907	4.205	4.449	4.655	4.832	4.989	
	.10	3.015	3.976	4.586	5.035	5.388	5.679	5.926	6.139	6.327	
	.05	3.927	5.040	5.757	6.287	6.707	7.053	7.347	7.60	7.826	
	.01	6.512	8.120	9.173	9.958	10.58	11.10	11.55	11.93	12.27	
5	.20	2.087	2.872	3.358	3.712	3.988	4.214	4.405	4.57	4.715	
	.10	2.850	3.717	4.264	4.664	4.979	5.238	5.458	5.648	5.816	
	.05	3.635	4.602	5.218	5.673	6.033	6.330	6.582	6.802	6.995	
	.01	5.702	6.976	7.804	8.421	8.913	9.321	9.669	9.972	10.24	
6	.20	2.036	2.788	3.252	3.588	3.850	4.065	4.246	4.403	4.540	
	.10	2.748	3.559	4.065	4.435	4.726	4.966	5.168	5.344	5.499	
	.05	3.461	4.339	4.896	5.305	5.628	5.895	6.122	6.319	6.493	
	.01	5.243	6.331	7.033	7.556	7.973	8.318	8.613	8.869	9.097	
7	.20	2.001	2.731	3.179	3.503	3.756	3.962	4.136	4.287	4.419	
	.10	2.680	3.451	3.931	4.28	4.555	4.780	4.972	5.137	5.283	
	.05	3.344	4.165	4.681	5.06	5.359	5.606	5.815	5.998	6.158	
	.01	4.949	5.919	6.543	7.005	7.373	7.679	7.939	8.166	8.368	
8	.20	1.976	2.689	3.126	3.440	3.686	3.886	4.055	4.201	4.330	
	.10	2.630	3.374	3.834	4.169	4.431	4.646	4.829	4.987	5.126	
	.05	3.261	4.041	4.529	4.886	5.167	5.399	5.597	5.767	5.918	
	.01	4.746	5.635	6.204	6.625	6.960	7.237	7.474	7.681	7.863	
9	.20	1.956	2.658	3.085	3.393	3.633	3.828	3.994	4.136	4.261	
	.10	2.592	3.316	3.761	4.084	4.337	4.545	4.721	4.873	5.007	
	.05	3.199	3.949	4.415	4.756	5.024	5.244	5.432	5.595	5.739	
	.01	4.596	5.428	5.957	6.348	6.658	6.915	7.134	7.325	7.495	
10	.20	1.941	2.632	3.053	3.355	3.590	3.782	3.944	4.084	4.206	
	.10	2.563	3.270	3.704	4.018	4.264	4.465	4.636	4.783	4.913	
	.05	3.151	3.877	4.327	4.654	4.912	5.124	5.305	5.461	5.599	
	.01	4.482	5.270	5.769	6.136	6.428	6.669	6.875	7.055	7.213	

Table A.8 (Continued)

v	α	t								
		2	3	4	5	6	7	8	9	10
(a) $t = 2\text{--}10$										
11	.20	1.928	2.612	3.027	3.325	3.557	3.745	3.905	4.042	4.162
	.10	2.540	3.234	3.658	3.965	4.205	4.401	4.568	4.711	4.838
	.05	3.113	3.820	4.256	4.574	4.823	5.028	5.202	5.353	5.487
	.01	4.392	5.146	5.621	5.970	6.247	6.476	6.672	6.842	6.992
12	.20	1.918	2.596	3.006	3.300	3.529	3.715	3.872	4.007	4.126
	.10	2.521	3.204	3.621	3.922	4.156	4.349	4.511	4.652	4.776
	.05	3.082	3.773	4.199	4.508	4.751	4.950	5.119	5.265	5.395
	.01	4.320	5.046	5.502	5.836	6.101	6.321	6.507	6.670	6.814
13	.20	1.910	2.582	2.988	3.279	3.505	3.689	3.844	3.978	4.095
	.10	2.505	3.179	3.589	3.885	4.116	4.305	4.464	4.602	4.724
	.05	3.055	3.735	4.151	4.453	4.690	4.885	5.049	5.192	5.318
	.01	4.260	4.964	5.404	5.727	5.981	6.192	6.372	6.528	6.667
14	.20	1.902	2.570	2.973	3.261	3.485	3.667	3.820	3.953	4.069
	.10	2.491	3.158	3.563	3.854	4.081	4.267	4.424	4.560	4.680
	.05	3.033	3.702	4.111	4.407	4.639	4.829	4.990	5.131	5.254
	.01	4.210	4.895	5.322	5.634	5.881	6.085	6.258	6.409	6.543
15	.20	1.896	2.560	2.960	3.246	3.467	3.648	3.800	3.931	4.046
	.10	2.479	3.140	3.540	3.828	4.052	4.235	4.390	4.524	4.641
	.05	3.014	3.674	4.076	4.367	4.595	4.782	4.940	5.077	5.198
	.01	4.168	4.836	5.252	5.556	5.796	5.994	6.162	6.309	6.439
16	.20	1.891	2.551	2.948	3.232	3.452	3.631	3.782	3.912	4.026
	.10	2.469	3.124	3.520	3.804	4.026	4.207	4.360	4.492	4.608
	.05	2.998	3.649	4.046	4.333	4.557	4.741	4.897	5.031	5.150
	.01	4.131	4.786	5.192	5.489	5.722	5.915	6.079	6.222	6.349
17	.20	1.886	2.543	2.938	3.220	3.439	3.617	3.766	3.895	4.008
	.10	2.460	3.110	3.503	3.784	4.004	4.183	4.334	4.464	4.579
	.05	2.984	3.628	4.020	4.303	4.524	4.705	4.858	4.991	5.108
	.01	4.099	4.742	5.140	5.430	5.659	5.847	6.007	6.147	6.270
18	.20	1.882	2.536	2.930	3.210	3.427	3.604	3.753	3.881	3.993
	.10	2.445	3.098	3.488	3.767	3.984	4.161	4.311	4.440	4.554
	.05	2.971	3.609	3.997	4.277	4.495	4.673	4.824	4.956	5.071
	.01	4.071	4.703	5.094	5.379	5.603	5.788	5.944	8.081	6.201
19	.20	1.878	2.530	2.922	3.200	3.416	3.592	3.740	3.867	3.979
	.10	2.445	3.087	3.474	3.751	3.966	4.142	4.290	4.418	4.531
	.05	2.960	3.593	3.977	4.253	4.469	4.645	4.794	4.924	5.038
	.01	4.046	4.670	5.054	5.334	5.554	5.735	5.889	6.022	6.141
20	.20	1.874	2.524	2.914	3.192	3.407	3.582	3.729	3.855	3.966
	.10	2.439	3.078	3.462	3.736	3.950	4.124	4.271	4.398	4.510
	.05	2.950	3.578	3.958	4.232	4.445	4.620	4.768	4.896	5.008
	.01	4.024	4.639	5.018	5.294	5.510	5.688	5.839	5.970	6.087

Table A.8 (Continued)

v	α	t								
		2	3	4	5	6	7	8	9	10
<i>(a) $t = 2\text{--}10$</i>										
24	.20	1.864	2.507	2.892	3.166	3.377	3.549	3.694	3.818	3.927
	.10	2.420	3.047	3.423	3.692	3.900	4.070	4.213	4.336	4.445
	.05	2.919	3.532	3.901	4.166	4.373	4.541	4.684	4.807	4.915
	.01	3.956	4.546	5.907	5.168	5.374	5.542	5.685	5.809	5.919
30	.20	1.853	2.490	2.870	3.140	3.348	3.517	3.659	3.781	3.887
	.10	2.400	3.017	3.386	3.648	3.851	4.016	4.155	4.275	4.381
	.05	2.888	3.486	3.845	4.102	4.302	4.464	4.602	4.720	4.824
	.01	3.889	4.455	4.799	5.048	5.242	5.401	5.536	5.653	5.756
40	.20	1.843	2.473	2.848	3.114	3.318	3.484	3.624	3.743	3.848
	.10	2.381	2.988	3.349	3.605	3.803	3.963	4.099	4.215	4.317
	.05	2.858	3.442	3.791	4.039	4.232	4.389	4.521	4.635	4.735
	.01	3.825	4.367	4.696	4.931	5.114	5.265	5.392	5.502	5.599
60	.20	1.833	2.456	2.826	3.089	3.290	3.452	3.589	3.707	3.809
	.10	2.363	2.959	3.312	3.562	3.755	3.911	4.042	4.155	4.254
	.05	2.829	3.399	3.737	3.977	4.163	4.314	4.441	4.550	4.646
	.01	3.762	4.282	4.595	4.818	4.991	5.133	5.253	5.356	5.447
120	.20	1.822	2.440	2.805	3.063	3.260	3.420	3.554	3.669	3.770
	.10	2.344	2.930	3.276	3.520	3.707	3.859	4.987	4.096	4.191
	.05	2.800	3.356	3.685	3.917	4.096	4.241	4.363	4.468	4.560
	.01	3.702	4.200	4.497	4.709	4.872	5.005	5.118	5.214	5.299
∞	.20	1.812	2.424	2.784	3.037	3.232	3.389	3.520	3.632	3.730
	.10	2.326	2.902	3.240	3.478	3.661	3.808	3.931	4.037	4.129
	.05	2.772	3.314	3.633	3.858	4.030	4.170	4.286	4.387	4.474
	.01	3.643	4.120	4.403	4.603	4.757	4.882	4.987	5.078	5.157

Table A.8 (Continued)

v	α	t								
		11	12	13	14	15	16	17	18	19
(b) $t=11-19$										
10	.20	4.316	4.414	4.503	4.585	4.660	4.730	4.795	4.856	4.913
	.10	5.029	5.134	5.229	5.317	5.397	5.472	5.542	5.607	5.668
	.05	5.722	5.833	5.935	6.028	6.114	6.194	6.269	6.339	6.405
	.01	7.356	7.485	7.603	7.712	7.812	7.906	7.993	8.076	8.153
11	.20	4.270	4.366	4.454	4.534	4.608	4.677	4.741	4.801	4.857
	.10	4.951	5.053	5.146	5.231	5.309	5.382	5.450	5.514	5.573
	.05	5.605	5.713	5.811	5.901	5.984	6.062	6.134	6.202	6.265
	.01	7.128	7.250	7.362	7.465	7.560	7.649	7.732	7.809	7.883
12	.20	4.231	4.327	4.413	4.492	4.565	4.633	4.696	4.755	4.810
	.10	4.886	4.986	5.077	5.160	5.236	5.308	5.374	5.436	5.495
	.05	5.511	5.615	5.710	5.798	5.878	5.953	6.023	6.089	6.151
	.01	6.943	7.060	7.167	7.265	7.356	7.441	7.520	7.594	7.665
13	.20	4.199	4.293	4.379	4.457	4.529	4.596	4.658	4.716	4.770
	.10	4.832	4.930	5.019	5.100	5.176	5.245	5.311	5.372	5.429
	.05	5.431	5.533	5.625	5.711	5.789	5.862	5.931	5.995	6.055
	.01	6.791	6.903	7.006	7.101	7.188	7.269	7.345	7.417	7.485
14	.20	4.172	4.265	4.349	4.426	4.498	4.564	4.625	4.683	4.737
	.10	4.786	4.882	4.970	5.050	5.124	5.192	5.256	5.316	5.373
	.05	5.364	5.463	5.554	5.637	5.714	5.786	5.852	5.915	5.974
	.01	6.664	6.772	6.871	6.962	7.047	7.126	7.199	7.268	7.333
15	.20	4.148	4.240	4.324	4.400	4.471	4.536	4.597	4.654	4.707
	.10	4.746	4.841	4.927	5.006	5.079	5.147	5.209	5.269	5.324
	.05	5.306	5.404	5.493	5.574	5.649	5.720	5.785	5.846	5.904
	.01	6.555	6.660	6.757	6.845	6.927	7.003	7.074	7.142	7.204
16	.20	4.127	4.218	4.301	4.377	4.447	4.512	4.572	4.628	4.681
	.10	4.712	4.805	4.890	4.968	5.040	5.107	5.169	5.227	5.282
	.05	5.256	5.352	5.439	5.520	5.593	5.662	5.727	5.786	5.843
	.01	6.562	6.564	6.658	6.744	6.823	6.898	6.967	7.032	7.093
17	.20	4.109	4.199	4.282	4.357	4.426	4.490	4.550	4.606	4.659
	.10	4.682	4.774	4.858	4.935	5.005	5.071	5.133	5.190	5.244
	.05	5.212	5.307	5.392	5.471	5.544	5.612	5.675	5.734	5.790
	.01	6.381	6.480	6.572	6.656	6.734	6.806	6.873	6.937	6.997
18	.20	4.093	4.182	4.264	4.339	4.407	4.471	4.531	4.586	4.638
	.10	4.655	4.746	4.829	4.905	4.975	5.040	5.101	5.158	5.211
	.05	5.174	5.267	5.352	5.429	5.501	5.568	5.630	5.688	5.743
	.01	6.310	6.407	6.497	6.579	6.655	6.725	6.792	6.854	6.912
19	.20	4.078	4.167	4.248	4.323	4.391	4.454	4.513	4.569	4.620
	.10	4.631	4.721	4.803	4.879	4.948	5.012	5.073	5.129	5.182
	.05	5.140	5.231	5.315	5.391	5.462	5.528	5.589	5.647	5.701
	.01	6.247	6.342	6.430	6.510	6.585	6.654	6.719	6.780	6.837

Table A.8 (Continued)

v	α	t								
		11	12	13	14	15	16	17	18	19
<i>(b) $t = 11\text{--}19$</i>										
20	.20	4.065	4.154	4.234	4.308	4.376	4.439	4.498	4.552	4.604
	.10	4.609	4.699	4.780	4.855	4.924	4.987	5.047	5.103	5.155
	.05	5.108	5.199	5.282	5.357	5.427	5.493	5.553	5.610	5.663
	.01	6.191	6.285	6.371	6.450	6.523	6.591	6.654	6.714	6.771
24	.20	4.024	4.111	4.190	4.262	4.329	4.391	4.448	4.502	4.552
	.10	4.541	4.628	4.708	4.780	4.847	4.909	4.966	5.021	5.071
	.05	5.012	5.099	5.179	5.251	5.319	5.381	5.439	5.494	5.545
	.01	6.017	6.106	6.186	6.261	6.330	6.394	6.453	6.510	6.563
30	.20	3.982	4.068	4.145	4.216	4.281	4.342	4.398	4.451	4.500
	.10	4.474	4.559	4.635	4.706	4.770	4.830	4.886	4.939	4.988
	.05	4.917	5.001	5.077	5.147	5.211	5.271	5.327	5.379	5.429
	.01	5.849	5.932	6.008	6.078	6.143	6.203	6.259	6.311	6.361
40	.20	3.941	4.025	4.101	4.170	4.234	4.293	4.348	4.399	4.447
	.10	4.408	4.490	4.564	4.632	4.695	4.752	4.807	4.857	4.905
	.05	4.824	4.904	4.977	5.044	5.106	5.163	5.216	5.266	5.313
	.01	5.686	5.764	5.835	5.900	5.961	6.017	6.069	6.119	6.165
60	.20	3.900	3.982	4.056	4.124	4.186	4.244	4.297	4.347	4.395
	.10	4.342	4.421	4.493	4.558	4.619	4.675	4.727	4.775	4.821
	.05	4.732	4.808	4.878	4.942	5.001	5.056	5.107	5.154	5.199
	.01	5.528	5.601	5.667	5.728	5.785	5.837	5.886	5.931	5.974
120	.20	3.859	3.938	4.011	4.077	4.138	4.194	4.246	4.295	4.341
	.10	4.276	4.353	4.422	4.485	4.543	4.597	4.647	4.694	4.738
	.05	4.641	4.714	4.781	4.842	4.898	4.950	4.998	5.044	5.086
	.01	5.375	5.443	5.505	5.562	5.614	5.662	5.708	5.750	5.790
∞	.20	3.817	3.895	3.966	4.030	4.089	4.144	4.195	4.242	4.287
	.10	4.211	4.285	4.351	4.412	4.468	4.519	4.568	4.612	4.654
	.05	4.552	4.622	4.685	4.743	4.796	4.845	4.891	4.934	4.974
	.01	5.227	5.290	5.348	5.400	5.448	5.493	5.535	5.574	5.611

Table A.8 (Continued)

ν	α	t								
		20	22	24	26	28	30	32	34	36
<i>(c) $t = 20\text{--}36$</i>										
19	.20	4.669	4.759	4.840	4.914	4.981	5.044	5.102	5.156	5.206
	.10	5.232	5.324	5.407	5.483	5.552	5.616	5.676	5.732	5.784
	.05	5.752	5.846	5.932	6.009	6.081	6.147	6.209	6.267	6.321
	.01	6.891	6.992	7.082	7.166	7.242	7.313	7.379	7.440	7.498
20	.20	4.652	4.742	4.822	4.895	4.963	5.025	5.082	5.136	5.186
	.10	5.205	5.296	5.378	5.453	5.522	5.586	5.645	5.700	5.752
	.05	5.714	5.807	5.891	5.968	6.039	6.104	6.165	6.222	6.275
	.01	6.823	6.922	7.011	7.092	7.168	7.237	7.302	7.362	7.419
24	.20	4.599	4.687	4.766	4.838	4.904	4.964	5.021	5.073	5.122
	.10	5.119	5.208	5.287	5.360	5.427	5.489	5.546	5.600	5.650
	.05	5.594	5.683	5.764	5.838	5.906	5.968	6.027	6.081	6.132
	.01	6.612	6.705	6.789	6.865	6.936	7.001	7.062	7.119	7.173
30	.20	4.546	4.632	4.710	4.779	4.844	4.903	4.958	5.010	5.058
	.10	5.034	5.120	5.197	5.267	5.332	5.392	5.447	5.499	5.547
	.05	5.475	5.561	5.638	5.709	5.774	5.833	5.889	5.941	5.990
	.01	6.407	6.494	6.572	6.644	6.710	6.772	6.828	6.881	6.932
40	.20	4.493	4.576	4.652	4.720	4.783	4.841	4.895	4.945	4.993
	.10	4.949	5.032	5.107	5.174	5.236	5.294	5.347	5.397	5.444
	.05	5.358	5.439	5.513	5.581	5.642	5.700	5.753	5.803	5.849
	.01	6.209	6.289	6.362	6.429	6.490	6.547	6.600	6.650	6.697
60	.20	4.439	4.520	4.594	4.661	4.722	4.778	4.831	4.880	4.925
	.10	4.864	4.944	5.015	5.081	5.141	5.196	5.247	5.295	5.340
	.05	5.241	5.319	5.389	5.453	5.512	5.566	5.617	5.664	5.708
	.01	6.015	6.090	6.158	6.220	6.277	6.330	6.378	6.424	6.467
120	.20	4.384	4.463	4.535	4.600	4.659	4.714	4.765	4.812	4.857
	.10	4.779	4.856	4.924	4.987	5.044	5.097	5.146	5.192	5.235
	.05	5.126	5.200	5.266	5.327	5.382	5.434	5.481	5.526	5.568
	.01	5.827	5.897	5.959	6.016	6.069	6.117	6.162	6.204	6.244
∞	.20	4.329	4.405	4.475	4.537	4.595	4.648	4.697	4.743	4.786
	.10	4.694	4.767	4.832	4.892	4.947	4.997	5.044	5.087	5.128
	.05	5.012	5.081	5.144	5.201	5.253	5.301	5.346	5.388	5.427
	.01	5.645	5.709	5.766	5.818	5.866	5.911	5.952	5.990	6.026

Table A.8 (*Continued*)

v	α	t							
		38	40	50	60	70	80	90	100
(d) $t = 38\text{--}100$									
30	.20	5.103	5.146	5.329	5.475	5.597	5.701	5.791	5.871
	.10	5.593	5.636	5.821	5.969	6.093	6.198	6.291	6.372
	.05	6.037	6.080	6.267	6.417	6.543	6.650	6.744	6.827
	.01	6.978	7.023	7.215	7.370	7.500	7.611	7.709	7.796
40	.20	5.037	5.078	5.257	5.399	5.518	5.619	5.708	5.786
	.10	5.488	5.529	5.708	5.850	5.969	6.071	6.160	6.238
	.05	5.893	5.934	6.112	6.255	6.375	6.477	6.566	6.645
	.01	6.740	6.782	6.960	7.104	7.225	7.328	7.419	7.500
60	.20	4.969	5.009	5.183	5.321	5.437	5.535	5.621	5.697
	.10	5.382	5.422	5.593	5.730	5.844	5.941	6.026	6.102
	.05	5.750	5.789	5.958	6.093	6.206	6.303	6.387	6.462
	.01	6.507	6.546	6.710	6.843	6.954	7.050	7.133	7.207
120	.20	4.899	4.938	5.106	5.240	5.352	5.447	5.530	5.603
	.10	5.275	5.313	5.476	5.606	5.715	5.808	5.888	5.960
	.05	5.607	5.644	5.802	5.929	6.035	6.126	6.205	6.275
	.01	6.281	6.316	6.467	6.588	6.689	6.776	6.852	6.919
∞	.20	4.826	4.864	5.026	5.155	5.262	5.353	5.433	5.503
	.10	5.166	5.202	5.357	5.480	5.582	5.669	5.745	5.812
	.05	5.463	5.498	5.646	5.764	5.863	5.947	6.020	6.085
	.01	6.060	6.092	6.228	6.338	6.429	6.507	6.575	6.636

Source: Values were extracted by permission from H. L. Harter, 1969, *Order Statistics and Their Use in Testing and Estimation*, Vol. 1, Aerospace Research Laboratory, USAF, U.S. Government Printing Office, Washington, D.C., pp. 648–657.

Table A.9 Upper Quantiles of the Dunnett's *t* Distribution: One-Sided Comparisons with Control

<i>v</i>	α	<i>m</i>								
		1	2	3	4	5	6	7	8	9
5	.05	2.02	2.44	2.68	2.85	2.98	3.08	3.16	3.24	3.30
	.01	3.37	3.90	4.21	4.43	4.60	4.73	4.85	4.94	5.03
6	.05	1.94	2.34	2.56	2.71	2.83	2.92	3.00	3.07	3.12
	.01	3.14	3.61	3.88	4.07	4.21	4.33	4.43	4.51	4.59
7	.05	1.89	2.27	2.48	2.62	2.73	2.82	2.89	2.95	3.01
	.01	3.00	3.42	3.66	3.83	3.96	4.07	4.15	4.23	4.30
8	.05	1.86	2.22	2.42	2.55	2.66	2.74	2.81	2.87	2.92
	.01	2.90	3.29	3.51	3.67	3.79	3.88	3.96	4.03	4.09
9	.05	1.83	2.18	2.37	2.50	2.60	2.68	2.75	2.81	2.86
	.01	2.82	3.19	3.40	3.55	3.66	3.75	3.82	3.89	3.94
10	.05	1.81	2.15	2.34	2.47	2.56	2.64	2.70	2.76	2.81
	.01	2.76	3.11	3.31	3.45	3.56	3.64	3.71	3.78	3.83
11	.05	1.80	2.13	2.31	2.44	2.53	2.60	2.67	2.72	2.77
	.01	2.72	3.06	3.25	3.38	3.48	3.56	3.63	3.69	3.74
12	.05	1.78	2.11	2.29	2.41	2.50	2.58	2.64	2.69	2.74
	.01	2.68	3.01	3.19	3.32	3.42	3.50	3.56	3.62	3.67
13	.05	1.77	2.09	2.27	2.39	2.48	2.55	2.61	2.66	2.71
	.01	2.65	2.97	3.15	3.27	3.37	3.44	3.51	3.56	3.61
14	.05	1.76	2.08	2.25	2.37	2.46	2.53	2.59	2.64	2.69
	.01	2.62	2.94	3.11	3.23	3.32	3.40	3.46	3.51	3.56
15	.05	1.75	2.07	2.24	2.36	2.44	2.51	2.57	2.62	2.67
	.01	2.60	2.91	3.08	3.20	3.29	3.36	3.42	3.47	3.52
16	.05	1.75	2.06	2.23	2.34	2.43	2.50	2.56	2.61	2.65
	.01	2.58	2.88	3.05	3.17	3.26	3.33	3.39	3.44	3.48
17	.05	1.74	2.05	2.22	2.33	2.42	2.49	2.54	2.59	2.64
	.01	2.57	2.86	3.03	3.14	3.23	3.30	3.36	3.41	3.45
18	.05	1.73	2.04	2.21	2.32	2.41	2.48	2.53	2.58	2.62
	.01	2.55	2.84	3.01	3.12	3.21	3.27	3.33	3.38	3.42
19	.05	1.73	2.03	2.20	2.31	2.40	2.47	2.52	2.57	2.61
	.01	2.54	2.83	2.99	3.10	3.18	3.25	3.31	3.36	3.40

Table A.9 (Continued)

<i>v</i>	α	<i>m</i>								
		1	2	3	4	5	6	7	8	9
20	.05	1.72	2.03	2.19	2.30	2.39	2.46	2.51	2.56	2.60
	.01	2.53	2.81	2.97	3.08	3.17	3.23	3.29	3.34	3.38
24	.05	1.71	2.01	2.17	2.28	2.36	2.43	2.48	2.53	2.57
	.01	2.49	2.77	2.92	3.03	3.11	3.17	3.22	3.27	3.31
30	.05	1.70	1.99	2.15	2.25	2.33	2.40	2.45	2.50	2.54
	.01	2.46	2.72	2.87	2.97	3.05	3.11	3.16	3.21	3.24
40	.05	1.68	1.97	2.13	2.23	2.31	2.37	2.42	2.47	2.51
	.01	2.42	2.68	2.82	2.92	2.99	3.05	3.10	3.14	3.18
60	.05	1.67	1.95	2.10	2.21	2.28	2.35	2.39	2.44	2.48
	.01	2.39	2.64	2.78	2.87	2.94	3.00	3.04	3.08	3.12
120	.05	1.66	1.93	2.08	2.18	2.26	2.32	2.37	2.41	2.45
	.01	2.36	2.60	2.73	2.82	2.89	2.94	2.99	3.03	3.06
∞	.05	1.64	1.92	2.06	2.16	2.23	2.29	2.34	2.38	2.42
	.01	2.33	2.56	2.68	2.77	2.84	2.89	2.93	2.97	3.00

Source: Values taken by permission from C. W. Dunnett, *J. Am. Stat. Assoc.*, **50**(1955):1115–1116.

APPENDIX B

SAS PROGRAMS

- Table B.1 SAS Program Used for Generation of Random Codes by Simple Randomization for Two Parallel Groups
- Table B.2 SAS Program Used for Generation of Random Codes by Simple Randomization for Three Parallel Groups
- Table B.3 SAS Program Used for Generation of Random Codes by Permuted-Block Randomization and Random Allocation for Two Parallel Groups
- Table B.4 SAS Program Used for Generation of Random Codes by Permuted-Block by Random Selection for Parallel Groups

Table B.1 SAS Program Used for Generation of Random Codes by Simple Randomization for Two Parallel Groups

```

options linesize =70 pagesize=40 nodate;
data one;
seed=4576891;
n=1;
p=0.5;
do center=1 to 4;
do sub=1 to 24;
tmt=ranbin(seed,n,p)+1;
subject=(14*100000)+(center*1000)+sub;
output;
end;
end;
run;
proc format;
value tmtf 1='Active Drug'
      2='Placebo';
value centerf 1='J. Smith, MD'
      2='M. Dole, MD '
      3='A. Hope, MD '
      4='C. Price, MD';
proc sort data=one; by center;
proc print data=one split='*' ; pageby center; by center;
id subject; var tmt;
label
      subject='Subject*Number*_____',
      tmt='Treatment *Assignment*_____';
format tmt tmtf. center centerf.;

title1 'Table 4.3.3' ;
title2 ' Example of Simple Randomization for Four Centers' ;
title3 ' Program:[DRUGXXX.PXXX014.SAS.NCKUJPL]SIMPRAN.SAS' ;
title4 ' Random Codes for Drug XXX, Protocol XXX-014' ;
title5 ' Double-blind, Randomized, Placebo Control, Two Parallel Groups' ;
run;
proc freq data=one;
tables center*tmt/nopercent nocol norow;
format tmt tmtf.;

run;

```

Table B.2 SAS Program Used for Generation of Random Codes by Simple Randomization for Three Parallel Groups

```

options linesize=70 pagesize=40 nodate;
data one;
seed=716891;
n=1;
p=1/3;
do center=1 to 4;
do sub=1 to 24;
seed=seed+(sub*(center-1));
tmt=rantbl(seed,p,p,p);
subject=(14*100000)+(center*1000)+sub;
output;
end;
end;
run;
proc print; run;
proc format;
value tmtf 1='Placebo'
      2='100 mg '
      3='200 mg ';
value centerf 1='J. Smith, MD'
      2='M. Dole, MD '
      3='A. Hope, MD '
      4='C. Price, MD';
proc sort data=one; by center;
proc print data=one split='*' ; pageby center; by center;
id subject; var tmt;
label
      subject='Subject*Number*_____',
      tmt='Treatment*Assignment*_____';
format tmt tmtf. center centerf.;

title1 ' Table 4.3.4 ' ;
title2 ' Example of Simple Randomization for Four Centers ' ;
title3 ' Program:[DRUGXXX.PXXX016.SAS.NCKUJPL]SIMPRAN.SAS ' ;
title4 ' Random Codes for Drug XXX, Protocol XXX-014 ' ;
title5 ' Double-blind, Randomized, Placebo Control, Three Parallel Groups ' ;
run;
proc freq data=one;
tables center*tmt/nopercent nocol norow;
format tmt tmtf.;

run;

```

Table B.3 SAS Program Used for Generation of Random Codes by Permuted-Block Randomization and Random Allocation for Two Parallel Groups

```

options linesize=70 pagesize=40 nodate;
proc plan ordered;
factors center=4 blocks=6 cell=4;
treatments t=4 random;
output out=perblock;
run;
data perblock; set perblock;
tmt=1;
if t gt 2 then tmt=2;
sub=_n_;
subject=(14*100000)+center*1000+(sub-(center-1)*24);
proc format;
value tmtf 1='Active Drug'
      2='Placebo';
value centerf 1='J. Smith, MD'
      2='M. Dole, MD'
      3='A. Hope, MD'
      4='C. Price, MD';
proc sort data=perblk; by center;
proc print data=perblock split='*' ; page by center; by center;
id subject;
var t tmt;
label
      subject='Subject*Number*_____',
      t=' Random*Permutation*_____',
      tmt='Treatment*Assignment*_____';
format tmt tmtf center centerf.;
title1 ' Table 4.3.7' ;
title2 ' Example of Permuted-Block Randomization' ;
title3 ' for Four Centers and a Block Size of Four' ;
title4 ' Program:[DRUGXXX.PXXX014.SAS.NCKUJPL]PERBLK.SAS' ;
title5 ' Random Codes for Drug XXX, Protocol XXX-014' ;
title6 ' Double-blind, Randomized, Placebo-Control, Two Parallel Groups' ;
run;
proc sort data=perblock; by center;
proc freq data=perblock;
tables center*tmt/nopercent nocol norow;
format tmt tmtf.; run;
proc freq data=perblock; by center;
tables blocks*tmt/nopercent nocol norow;
format tmt tmtf.; run;
proc plan ordered;
factors center=4 cell=24;
treatments t=24 random;

```

Table B.3 (*Continued*)

```

output out=perblock;
run;
data perblock;    set perblock;
tmt=1;
if t gt 12 then tmt=2;
sub=_n_;
subject=(14*100000)+center*1000+(sub-(center-1)*24);
proc format;
value tmtf 1='Active Drug'
      2='Placebo' ;
proc sort data=perblk;  by center;
proc print data=perblock split='*' page by center; by center;
id subject;
var   t tmt;
label
      subject='Subject*Number*_____',
      t=' Random *Permutation*_____',
      tmt='Treatment *Assignment*_____';
format tmt tmtf. center centerf.:
title1 '                               Table 4.3.6' ';
title2 ' Example of Random Allocation for Four Centers' ';
title3 ' Program:[DRUGXXX.PXXX014.SAS.NCKUJPL]RANALC.SAS' ';
title4 ' Random Codes for Drug XXX, Protocol XXX-014' ';
title5 ' Double-blind, Randomized, Placebo Control, Two Parallel Groups' ';
run;
proc freq data=perblock;
tables center*tmt/nopercent nocol norow;
format tmt tmtf.:;
run;

```

Table B.4 SAS Program Used for Generation of Random Codes by Permuted-Block by Random Selection for Parallel Groups

```

options linesize=70 pagesize=40 nodate;
data one;
do bloc=1 to 6;
do sub=1 to 4;
input tmt @;
output;
end; end;
cards;
1 1 2 2 2 2 1 1 1 2 1 2 1 2 2 1 2 1 2 1 2 1 1 2
data one; set one;
do center=1 to 4;
output;
end;
proc sort data=one; by center block sub;
proc plan ordered;
factors center=4 nblock=6;
treatments block=6 random;
output out=perblock;
run;
proc sort data=perblock; by center block ;
data perblock; merge perblock one; by center block ;
proc sort data=perblock; by center nblock sub;
proc print;
data perblock; set perblock;
sub=_n_;
subject=(14*100000)+center*1000+(sub-(center-1)*24);
proc format;
value tmtf 1='Active Drug'
      2='Placebo    ';
value centerf 1='J. Smith, MD'
      2='M. Dole, MD '
      3='A. Hope, MD '
      4='C. Price, MD'
proc sort data=perblk; by center;
proc print data=perblock split='*' ; pageby center; by center;
id subject;
var tmt;
label subject='Subject*Number *_____',
      tmt="Treatment*Assignment*_____";
format tmt tmtf. center centerf.;
```

Table B.4 (*Continued*)

```
title1 ' Table 4.3.8      ';
title2 ' Example of Permutated-Block Randomization by Random Selection';
title3 ' of Blocks for Four Centers and a Block Size of Four';
title4 ' Program:[DRUGXXX.PXXX014.SAS.NCKUJPL]PERBLK.SAS';
title5 ' Random Codes for Drug XXX, Protocol XXX-014';
title6 ' Double-blind, Randomized, Placebo-Control, Two Parallel Groups';
run;
proc sort data=perblock; by center;
proc freq data=perblock;
tables center*tmt/nopercent nocol norow;
format tmt tmtf.; run;
proc freq data=perblock; by center;
tables nblock*tmt/nopercent nocol norow;
format tmt tmtf.; run;
```

BIBLIOGRAPHY

- Agresti, A. (1983). Testing marginal homogeneity for ordinal categorical variables. *Biometrics*, **39**, 505–510.
- Agresti, A. (1984). *Analysis of Ordered Categorical Data*. Wiley, New York.
- Agresti, A. (1989). A survey of models for repeated ordered categorical response data. *Stat. Med.*, **8**, 1209–1224.
- Agresti, A. (2002). *Categorical Data Analysis*, 2nd Ed. Wiley, New York.
- Agresti, A., and Wackerly, D. (1977). Some exact conditional tests of independence for $R \times C$ cross-classification tables. *Psychometrika*, **42**, 111–125.
- Ahn, C. (1998). An evaluation of phase I cancer clinical trial design. *Stat. Med.*, **7**, 1537–1549.
- Albanse, M. A., Clarke, W. R., Adams, H. P., Woolson, R. F., and TOAST Investigators (1994). Ensuring reliability of outcome measures in multicenter clinical trials of treatments for acute ischemic stroke. *Stroke*, **25**, 1746–1751.
- Altman, D., and Bland, M. (1995). Absence of evidence is not evidence of absence. *Br. Med. J.*, **311**, 485.
- Amberson, J. B., McMahon, B. T., and Pinner, M. A. (1931). A clinical trial of sanocrysin in pulmonary tuberculosis. *Amer. Rev. Tuber.*, **24**, 401–435.
- Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Process*. Springer, New York.
- Andersen, P. K., and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Stat.*, **10**, 1100–1120.
- Angell, M., and Kassire, J. P. (1994). Setting the record straight in the breast cancer trials (editorial). *New Engl. J. Med.*, **330**, 1448–1450.
- Anonymous (2001). Retraction. *Journal of Clinical Oncology*, **19**, 2973.
- APA (1987). *Diagnostic and Statistical Manual of Mental Disorders*, 3d Ed. American Psychiatric Association, Washington, DC.
- Armitage, P. (1975). *Sequential Medical Trials*, 2d Ed. Blackwell Scientific, Oxford.
- Armitage, P., and Berry, G. (1987). *Statistical Methods in Medical Research*. Blackwell Scientific Publications, Oxford.
- Armitage, P., McPherson, C. K., and Rowe, B. C. (1969). Repeated significance tests on accumulating data. *J. Roy. Stat. Soc., A* **132**, 235–244.
- Attal, M., and the Intergroupe Francis Du Myelome (1996). A prospective, randomized trial of autologous bone marrow transplantation and chemotherapy in multiple myeloma. *New Engl. J. Med.*, **335**, 91–97.
- Bailar, J. C. (1992). Some uses of statistical thinking. In *Medical Uses of Statistics*, Ed. by Bailar, J. C., and Mosteller, F. New England Journal of Medicine Books, Boston, 5–26.
- Bailar, J. C., and Mosteller, F. (1986). *Medical Uses of Statistics*. Massachusetts Medical Society, Waltham, MA.
- Bailey, C. (1992). Biguanides and NIDDM. *Diabet. Care*, **15**, 755–772.

- Balaam, L. N. (1968). A two-period design with t^2 experimental units. *Biometrics*, **24**, 61–73.
- Barry, M. J., Fowler, F. J., Jr., O'Leary, M. P., Bruskewitz, R. C., Holtgrewe, H. L., Mebust, W. K., and Cockett, A. T. (1992). The American Urological Association Symptom Index for Benign Prostatic Hyperplasia. *J. Urol.*, **148**, 1549–1557.
- Barst, R. J., Rubin, L. J., Long, W. A., McGoon, M. D., Rich, S., Badesch, D. B., Groves, B. M., Tapson, V. F., Bourge, R. C., Brundage, B. H., Koerner, S. K., Langgleben, D., Keller, C. A., Murali, S., Uretsky, B. F., Clayton, L. M., Jobsis, M. M., Blackburn, S. D., Shortino, D., and Crow, J. W. for the Primary Pulmonary Hypertension Study Group (1996). A comparison of continuous intravenous epoprostenol (prostacyclin) with conventional therapy for primary pulmonary hypertension. *New Engl. J. Med.*, **334**, 296–301.
- Bazell, R. (1998). *Her-2: The making of Herceptin®, a revolutionary treatment for breast cancer*. Random House, New York.
- Begg, C. B. (1990). On inference for Wei's biased coin design on the randomized play-by-winner rule. *Biometrika*, **76**, 467–484.
- Begg, C., Cho, M., Eastwood, S., et al. (1996). Improving the quality of reporting of randomized controlled trials, the CONSORT statement. *J. Am. Med. Assoc.*, **276**, 637–639.
- Berger, R. L. (1982). Multiparameter hypothesis testing in acceptance sampling. *Technometr.*, **24**, 295–300.
- Berger, R. L., and Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests, and equivalence confidence sets. *Stat. Sci.*, **11**, 283–319.
- Berger, R. L., and Sidik, K. (2003) Exact unconditional tests for 3×3 matched-pairs design, *Statistical Methods in Medical Research*, **12**, 91–108.
- Berry, D. A. (1990). *Statistical Methodology in the Pharmaceutical Science*. Dekker, New York.
- Beta-Blocker Heart Attack Trial Research Group (1982). A randomized trial of propranolol in patients with acute myocardial infarction I. Mortality results. Beta-blocker Heart Attack Trial—Final Report. *J. Am. Med. Assoc.*, **247**, 1707–1714.
- Beta-Blocker Heart Attack Trial Research Group (1983). A randomized trial of propranolol in patients with acute myocardial infarction: Morbidity. *J. Am. Med. Assoc.*, **250**, 2814–2819.
- Bhapkar, V. P. (1966). A note on the equivalence of two criteria for hypothesis in categorical data. *J. Am. Stat. Assoc.*, **61**, 256–264.
- Bivens, L. V., and Macfarlane, D. K. (1994). Fraud in breast cancer trials. *New Engl. J. Med.*, **330**, 1461.
- Blackwelder, W. C. (1982). "Proving the null hypothesis" in clinical trials. *Controlled Clin. Trials*, **3**, 345–353.
- Blackwell, D., and Hodges, J. L., Jr., (1957) Design for the control of selection bias. *Ann. Math. Stat.*, **28**, 449–460.
- Bloomfield, S. S. (1969). Conducting the clinical drug study. In *Proceedings of the Institute on Drug Literature Evaluation*. American Society of Hospital Pharmacists, Washington, DC, 147–154.
- Blyth, C. R., and Still, H. A. (1983). Binomial confidence intervals. *J. Am. Stat. Assoc.*, **78**, 108–116.
- BMS (1994). Dose-response study of various dose levels of metformin hydrochloride compared to placebo in patients with non-insulin dependent diabetes mellitus. Bristol-Myers Squibb, CV 138-001, Princeton, NJ.
- Box, G. E. P., and Draper, N. R. (1987). *Empirical Model-Building and Response Surface*. Wiley, New York.
- Box, G. E. P., Hunter, J. S., and Hunter, W. G. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. Wiley, New York.
- Boyarsky, S. J., and Paulson, D. F. (1977). A new look at bladder neck obstruction by the Food and Drug Administration regulators: Guideline for investigation of benign prostatic hypertrophy. *Trans. Am. Assoc. Genito. Surg.*, **68**, 29–32.

- Boylston, A. W. (2002). Clinical investigation of smallpox in 1767. *New Engl. J. Med.*, **346**, 1326–1328.
- Breslow, N. E., and Day, N. E. (1980). *Statistical Methods in Cancer Research. Vol. 1: The Analysis of Case-Control Studies*. Oxford University Press, New York.
- Breslow, N. E., and Day, N. E. (1987). *Statistical Methods in Cancer Research. Vol. 2: The Analysis of Cohort Studies*. Oxford University Press, New York.
- Broder, S. (1994). Fraud in breast cancer trials. *New Engl. J. Med.*, **330**, 1460–1461.
- Brody, H. (1981). *Placebos and the philosophy of Medicine*. University of Chicago Press, Chicago.
- Brody, H. (1982). The lie that heals: The ethics of giving placebos. *Ann. Int. Med.*, **97**, 112–118.
- Brown, B. W. (1980). The crossover experiment for clinical trials. *Biometrics*, **36**, 69–79.
- Brown, L. D., Hwang, J. T. G., and Munk, A. (1997). An unbiased test for bioequivalence problem. *Ann. Stat.*, **25**, 2345–2367.
- Brunelle, R., and Wilson, M. (1996). Analysis and interpretation of standard laboratory values. Presented at Biopharmaceutical Section Workshop on Adverse Events, Bethesda, MD, October 28, 1996.
- Buncher, C. R., and Tsay, J. Y. (1994). *Statistics in Pharmaceutical Industry*, 2d Ed. Dekker, New York.
- Buyse, M. E., Staquet, M. J., and Sylvester, R. J. (1984). *Cancer Clinical Trials: Methods and Practice*. Oxford Medical Publications, New York.
- Byar, D. P., and Piantadosi, S. (1985). Factorial designs for randomized clinical trials. *Cancer Treat. Rep.*, **69**, 1055–1063.
- Byar, D. P., Simon, R. M., Friedwald, W. T., Schlesselman, J. J., DeMets, D. L., Ellenberg, J. H., Gail, M. L. and Ware, J. H. (1976). Randomized clinical trials—Perspectives on some recent ideas. *New Engl. J. Med.*, **295**, 74–80.
- Byington, R. P., Curb, J. D., and Mattson, M. E. (1985). Assessment of double-blindness at the conclusion of the beta-blocker heart attack trial. *J. Am. Med. Assoc.*, **263**, 1733–1783.
- Calimlim, J. F., Wardell, W. M., Lasagna, L., and Gillies, A. J. (1977). Analgesic efficacy of an orally administered combination of pentazocine and aspirin, with observations on the use and statistical efficiency of “global” subjective efficacy ratings. *Clin. Pharmacol. Ther.*, **21**, 34–43.
- Cancer Therapy Evaluation Program, the US National Cancer Institutes (1999). *Common Toxicity Criteria Manual*, Bethesda, MD.
- Cancer Therapy Evaluation Program, the US National Cancer Institutes (2001). *NCI Guidelines: Expedited Adverse Event Reporting Requirements for NCI Investigational Agents*, Bethesda, MD.
- Cancer Therapy Evaluation Program, the US National Cancer Institutes (2002). *Clinical Data Update System (CDUS) v3.0, Notice of Modifications*, Bethesda, MD.
- Canner, P. L. (1981). Practical aspects of decision-making in clinical trials: The coronary drug project as a case study. *Controlled Clin. Trials*, **1**, 363–376.
- Capizzi, T., and Zhang, J. (1996). Testing the hypothesis that matters for multiple primary endpoints. *Drug Info. J.*, **30**, 949–956.
- Carey, R. A., Eby, R. Z., Beg, M. A., McNally, C. F., and Fox, M. J. (1984). Patient selection of blood pressure in clinical trials. *Clin. Ther.*, **7**, 121–126.
- CAST (1989). Cardiac Arrhythmia Suppression Trial. Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *New Engl. J. Med.*, **321**, 406–412.
- CBER/FDA Memorandum (1999). Summary of CBER considerations on selected aspects of active controlled trial design and analysis for the evaluation of thrombolytics in acute MI. Center for Biological Evaluation and Research/Food and Drug Administration Rockville, Maryland, U.S.A.
- Chakravorti, S. R., and Grizzle, J. E. (1975). Analysis of data from multiclinics experiments. *Biometrics*, **31**, 325–338.

- Chalmers, T. C., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., and Ambroz, A. (1981). A method for assessing the quality of a randomized controlled trial. *Controlled Clin. Trials*, **1**, 31–49.
- Chan, J. C., Tomlinson, B., Critchley, J. A., Cockram, C. S., and Walden, R. (1993). Metabolic and hemodynamic effects of Metformin and Glibenclamide in normotensive NIDDM patients. *Diabet. Care*, **16**, 1035–1038.
- Chan, I. S. F. (1998). Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. *Stat. Med.*, **17**, 1403–1413.
- Chan, I. S. F. (2003). Proving non-inferiority or equivalence of two treatments with dichotomous endpoints using exact methods. *Stat. Meth. Med. Res.* Vol. **12**, 37–58.
- Chan, I. S. F., and Bohidar, N. R. (1998). Exact power and sample size for vaccine efficacy studies. *Comm. Stat. A—Theory Meth.*, **27**, 1305–1322.
- Chan, I. S. F., and Zhang, Z. (1999). Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics*, **55**, 1202–1209.
- Chan, I. S. F., Wang, W. W. B., and Heyse, J. F. (2003). Vaccine Clinical Trials. In *Encyclopedia of Biopharmaceutical Statistics*, 2nd Ed. Ed. by Chow, S.C. Dekker, New York, 1005–1022.
- Chang, M. N. (1989). Confidence intervals for a normal mean following group sequential test. *Biometrics*, **45**, 247–254.
- Chang, M. N., and O'Brien, P. C. (1986). Confidence intervals following group sequential test. *Controlled Clin. Trials*, **7**, 18–26.
- Chang, M. N., Therneau, T. M., Wieand, H. S., and Cha, S. S. (1987). Designs for group sequential phase II clinical trials. *Biometrics*, **43**, 865–874.
- Chang, M. N., Wieand, H. S., and Chang, V. T. (1989). The bias of the sample proportion following a group sequential phase II trial. *Stat. Med.*, **8**, 563–570.
- Chang, M. N., Guess, H. A., and Heyse, J. F. (1994). Reduction in burden of illness: A new efficacy measure for prevention trials. *Stat. Med.*, **13**, 1807–1814.
- Chang, M. S. (2000). A two-sample comparison for multiple ordered event data. *Biometrics*, **56**, 183–189.
- Chang, M. S., and Wang, M. C. (1999). Conditional regression analysis for recurrence time data. *The J. Am. Stat. Assoc.*, **94**, 1221–1230.
- Chen, M.G. (2003). Interactive Voice Randomization System (IVRS) in the *Encyclopedia of Biopharmaceutical Statistics*, 2nd Ed. Ed. by Chow, S. C. Dekker, Inc., New York.
- Chen, J. J., Tsong, T., and Kang, S-h. (2000). Tests for equivalence and non-inferiority between two proportions. *Drug Info. J.*, **34**, 569–578.
- Chen, T. T. (1997). Optimal three-stage designs for phase II cancer clinical trials. *Stat. Med.*, **16**, 2701–2711.
- Chen, T. T., and Ng, T. H. (1998). Optimal flexible designs in phase II cancer clinical trials. *Stat. Med.*, **17**, 2301–2312.
- Cheng, B., and Chow, S.C. (2003). Validity of LOCF. In *Encyclopedia of Biopharmaceutical Statistics*, 2nd Ed. by Chow, S.C. Dekker, New York, 1023–1029.
- Chevret, S. (1993). The continual reassessment method in cancer phase I clinical trials: a simulation study. *Stat. Med.*, **12**, 1093–1108.
- Chinchilli, V. M., and Bortey, E. B. (1991). Testing for consistency in a single multi-center trial. *J. Biopharm. Stat.*, **1**, 67–80.
- Chinchilli, V. M., and Esinhart, J. D. (1996). Design and analysis of intra-subject variability in cross-over design. *Stat. Med.*, **15**, 1619–1634.
- Cholesterol and Recurrent Events Trial Investigators (1996). The effect of Pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. *New Engl. J. Med.*, **335**, 1001–1009.

- Chow, S. C. (1996). Statistical considerations for replicated design. In *Bioavailability, Bioequivalence and Pharmacokinetics Studies, Proceedings of FIP Bio-International '96*, Ed. by Midha, K. K. and Nagai, T. Business Center for Academic Societies, Tokyo, Japan, 107–112.
- Chow, S. C. (1997). Good statistics practice in drug development and regulatory approval process. *Drug Info. J.*, **31**, 1157–1166.
- Chow, S. C. (1999). Individual bioequivalence – a review of the FDA draft guidance, *Drug Info. J.*, **33**, 435–444.
- Chow, S. C. (2003). *Encyclopedia of Biopharmaceutical Statistics*, 2nd Ed. Dekker, New York.
- Chow, S. C., and Liu, J. P. (2000). *Design and Analysis of Bioavailability and Bioequivalence Studies*, 2nd Ed., Revised and Expanded. Dekker, New York.
- Chow, S. C., and Liu, J. P. (1992). On assessment of bioequivalence under a higher-order crossover design. *J. Biopharm. Stat.*, **2**, 239–256.
- Chow, S. C., and Liu, J. P., (1995a). *Statistical Design and Analysis in Pharmaceutical Science*. Dekker, New York.
- Chow, S. C., and Liu, J. P. (1995b). Current issues in bioequivalence trials. *Drug Info. J.*, **29**, 795–804.
- Chow, S. C., and Liu, J. P. (1997). Meta-analysis for bioequivalence review. *J. Biopharm. Stat.*, **7**, 97–111.
- Chow, S. C. (2001). Bridging studies in clinical research. Presented at the *2001 Symposium on APEC Network of Pharmaceutical Regulatory Science—APEC Joint Research Project on Bridging Study*, May 23–26, Taipei, Taiwan.
- Chow, S. C., and Shao, J. (1997). Statistical methods for two-sequence, three-period crossover designs with incomplete data. *Stat. Med.*, **16**, 1031–1039.
- Chow, S. C., and Shao, J. (2002a). *Statistics in Drug Research—Methodologies and Recent Developments*. Dekker, New York.
- Chow, S. C., and Shao, J., (2002b). A note on statistical methods for therapeutic equivalence. *Controlled Clin. Trials*, **23**, 515–520.
- Chow, S. C., and Shao, J. (2003). Analysis of clinical data with breached blindness, *Stat. Med.* **22**, in press.
- Chow, S. C., Shao, J., and Hu, O. Y. P. (2002). Assessing sensitivity and similarity in bridging studies. *J. Biopharm. Stat.*, **12**, 385–400.
- Chow, S. C., Shao, J., and Ho, H. C. (2000). Statistical analysis for placebo-challenging design in clinical trials, *Stat. Med.* **19**, 1029–1037.
- Chow, S. C., Shao, J., and Wang, H. (2002). Individual bioequivalence testing under 2×3 designs, *Stat. Med.* **21**, 629–648.
- Chow, S. C., Shao, J., and Wang, H. (2003a). Statistical tests for population bioequivalence, *Statistica Sinica*, **13**, 539–554.
- Chow, S. C., Shao, J., and Wang, H. (2003b). *Sample Size Calculation in Clinical Research*. Dekker, New York.
- Chow, S. C., and Wang, H. (2001). On sample size calculation in bioequivalence trials, *J. Pharmacokinetics and Pharmacodynamics*, **28**, 155–169.
- Chuang, C. (1987). The analysis of a titration study. *Stat. Med.*, **6**, 583–590.
- Chuang-Stein, C. (1993). Personal communications. Upjohn Company, Kalamazoo, MI.
- Chuang-Stein, C., and Shih, W. J. (1991). A note of the analysis of titration studies. *Stat. Med.*, **10**, 323–328.
- Chuang-Stein, C. (1996). Summarizing laboratory data with different reference ranges in multi-center clinical trials. *Drug Info. J.*, **26**, 77–84.

- Clinical Trials Group Study 343 Team (1998). Maintenance anteretroviral therapies in HIV-infected subjects with undetectable plasma HIV RNA after triple-drug therapy. *New Engl. J. Med.*, **339**, 1261–1268.
- Clopper, C. J., and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**, 404–413.
- COBALT Investigators (1997). A comparison of continuous infusion of alteplase with double-bolus administration for acute myocardial infarction. *New Engl. J. Med.*, **337**, 1124–1130.
- Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, **10**, 101–129.
- Cochran, W. G. (1997). *Sampling Techniques*. Wiley, New York.
- Cochran, W. G., and Cox, G. M. (1957). *Experimental Designs*, 2nd Ed. Wiley, New York, p. 18.
- Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York.
- Collins, R., Peto, R., Baigent, C., and Sleight, P. (1997). Aspirin, heparin, and fibrinolytic therapy in selected acute myocardial infarction. *New Engl. J. Med.*, **336**, 847–860.
- Colton, T. (1974). *Statistics in Medicine*. Little Brown, Boston.
- Com-nougue, C., Rodary, C., and Patte, C. (1993). How to establish equivalence when data are censored: a randomized trial of treatments for B Non-Hodgkin lymphoma. *Stat. Med.*, **12**, 1353–1364.
- Congress (1993). Section 492B, National Institutes of Health Reauthorization Bill, the 103rd U.S. Congress.
- Coniff, R. F., Shapiro, J. A., Seaton, T. B., and Bray, G. A. (1995). Multicenter placebo-controlled trial comparing acarbose (Bay g 5421) with placebo, tolbutamide, and tolbutamide-plus-acarbose in non-insulin-dependent diabetes mellitus. *Am. J. Med.*, **98**, 443.
- Conover, W. J. (1980). *Practical Nonparametric Statistics*. Wiley, New York.
- Cooppan, R. (1994). Pathophysiology and current treatment of NIDDM. Presented at Diabetes Education Updates. Harvard Medical School, Boston, MA.
- Cornell, R. G. (1990). Handling dropouts and related issues. In *Statistical Methodology in the Pharmaceutical Sciences*, Ed. by Berry, D. A. Dekker, New York.
- Cornell, R. G., Landenberger, B. D., and Bartlett, R. H. (1986). Randomized play-by-winner clinical trials. *Comm. Stat. Theory Meth.*, **15**, 159–178.
- Cornfield, J. (1978). Randomization by group: a formal analysis. *Am. J. Epidemiol.*, **108**, 100–102.
- Coronary Drug Project Research Group (1973). The coronary drug project: Design, method, and baseline results. *Circul.*, **47** (suppl. I), I-1–I-50.
- Coronary Drug Project Research Group (1980). Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Project. *New Engl. J. Med.*, **303**, 1038–1041.
- Cox, D. R. (1952). A note of the sequential estimation of means. *Proc. Camb. Phil. Soc.*, **48**, 447–450.
- Cox, D. R. (1972). Regression models and life tables. *J. Royal Stat. Soc., Series B*, **34**, 187–220.
- Cox, D. R. (1990). Discussion of paper by C. B. Begg. *Biometrika*, **77**, 483–484.
- Cox, D. R., and Snell, E. J. (1989). *Analysis of Binary Data*. 2nd Ed. Chapman and Hall, New York.
- CPMP Working Party on Efficacy on Medicinal Products (1990). Good clinical practice for trials on medicinal products in the European Community. *Pharmacol. Toxicol.*, **67**, 361–372.
- CPMP Working Party on Efficacy on Medicinal Products (1994). Biostatistical methodology in clinical trials in applications for marketing authorisations for medicinal products, European Commission, Brussels, Belgium.
- CPMP Working Party on Efficacy on Medicinal Products (1995). Biostatistical methodology in clinical trials in applications for marketing authorisations for medicinal products: Note for guidance. *Stat. Med.*, **14**, 1659–1682.
- Cramer, J. A., Mattson, R. H., Prevey, M. L., Scheyer, R. D., and Ouellette, V. L. (1989). How often is medication taken as prescribed. *J. Am. Med. Assoc.*, **261**, 3273–3277.

- Conaway, M. R., and Petroni, G. R. (1996) Designs for phase II trials allowing for a trade-off between response and toxicity. *Biometrics*, Vol. 52, 1375–1386.
- Crawford, E. D., Eisenberg, M. A., Mcleod, D. G., Spaulding, J. T., Benson, R., Dorr, A., Blumenstein, B. A., Davis, M. A., and Goodman, P. J. (1989). A controlled trial of Leuprolide with and without flutamide in prostatic carcinoma. *New Engl. J. Med.*, **321**, 419–424.
- Crépeau, H., Koziol, J., Reid, N., and Yuh, Y. S. (1985). Analysis of incomplete multivariate data from repeated measurement experiments. *Biometrics*, **41**, 505–514.
- Crowder, M. J., and Hand, D. J. (1990). *Analysis of Repeated Measures*, Chapman and Hall, London.
- Culliton, B. J. (1983). Copying with fraud: The Darsee Case. *Science*, **220**, 31–35.
- D'Agostino, R. B. Sr. (2003). Special issue in non-inferiority trials. *Stat. Med.*, **22**, 163–336.
- Dannenberg, O., Dette, H., and Munk, A. (1994). An extension of Welch's approximation *t*-solution to comparative bioequivalence trials. *Biometrika*, **81**, 91–101.
- Davis, C. S., and Chung, Y. (1995). Randomization model methods for evaluating treatment efficacy in multicenter clinical trials. *Biometrics*, **51**, 1163–1174.
- Davis, R., Ribner, H. S., Keung, E., Sonnenblick, E. H., and LeJemtel, T. H. (1979). Treatment of chronic congestive heart failure with catopril, an oral inhibitor of angiotension-converting enzyme. *New Engl. J. Med.*, **301**, 117–121.
- Davis, K. L., Thal, L. J., Gamzu, E. R., Davis, C. S., Woolson, R. F., Gracon, S. I., Drachman, D. A., Schneider, L. S., Whitehouse, P. J., Hoover, T. M., Morris, J. C., Kawas, G. H., Knopman, D. S., Earl, N. L., Jumar, V., Doody, R. S., and the Tacrine Collaborative Study Group (1992). A double-blind, placebo-controlled multicenter study for Alzheimer's disease. *New Engl. J. Med.*, **327**, 1253–1259.
- Day, S., Fayers, P., and Harvery, D. (1998). Double data entry: what value, what price? *Controlled Clin. Trials*, **19**, 15–24.
- DeMets, D., and Lan, K. K. G. (1994). Interim analysis: The alpha spending function approach. *Stat. Med.*, **13**, 1341–1352.
- DeMets, D. L. (2000). Relationships between data monitoring committees. *Controlled Clin. Trials*, **21**, 54–55.
- Demetri G. D., von Mehren M., Blanke C. D., Van den Abbeele A. D., Eisenberg B., Roberts P. J., Heinrich M. C., Tuveson D. A., Singer S., Janicek M., Fletcher J. A., Silverman S. G., Silberman S. L., Capdeville R., Kiese B., Peng B., Dimitrijevic S., Druker B. J., Corless C., Fletcher C. D., and Joensuu H. (2002). Efficacy and safety of imatinib mesylate in advanced gastrointesinal stromal tumors. *New Engl. J. Med.*, **347**, 472–480.
- Demiroglu, H., Ozcebe, O. I., Barista, I., Dundar, S., and Eldem, B. (2000). Interferon alfa-2b, colchicines, and benzathine penicillin versus colchicines and benzathine penicillin in Behcet's disease. *Lancet*, **355**, 605–609.
- DerSimonian, R. and Laird, N. (1986) Meta-analysis in clinical trials, *Controlled Clin. Trials*, Vol. 7, 177–188.
- Devereaux, P. J., Manns, B. J., Ghali, W. A., Quan, H., and Guyatt, G. H. (2002). The reporting of methodological factors in randomized controlled trials and association with a journal policy to promote adherence to the Consolidated Standards of Reporting Trials (CONSORT) checklist. *Controlled Clin. Trials*, **23**, 380–388.
- Devereaux, P. J., Manns, B. J., Ghali, W. A., Quan, H., Lacchetti, C., Monotir, V. M., Bhandari, M., and Guyatt, G. H. (2001). Physician interpretations and textbooks definitions of blinding terminology in randomized controlled trials. *J. Am. Med. Assoc.*, **285**, 2000–2003.
- Deyo, R. A., Walsh, N. E., Schoenfeld, L. S., and Ramamurthy, S. (1990). Can trials of physical treatments be blinded: The example of transcutaneous electrical nerve stimulation for chronic pain. *Am. J. Phys. Med. Rehab.*, **69**, 6–10.

- Diabetes Prevention Program Research Group (1999). The Diabetes Prevention Program: Design and methods for a clinical trial in the prevention of type 2 diabetes. *Diabetes Care*, **22**, 623–634.
- Diabetes Prevention Program Research Group (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin, *New Engl. J. Med.*, **345**, 393–403.
- Dietrich, F. H., and Kearns, T. J. (1986). *Basic Statistics: An Inferential Approach*. Dellen, San Francisco.
- Diggle, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics*, **44**, 959–971.
- Diggle, P. J. (1989). Test for random dropouts in repeated measurement data. *Biometrics*, **45**, 1255–1258.
- Diggle, P. J., and Kenward, M. G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Appl. Stat.*, **43**, 49–93.
- Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford Science, New York.
- Dixon, D. O., and Lagakos, S. W. (2000). Should data and safety monitoring boards share confidential interim data. *Controlled Clin. Trials*, **21**, 1–6.
- Djulhegovic, B., and Clarke, M. (2001). Scientific and ethical issues in equivalence trials. *J. Am. Med. Assoc.*, **285**, 1206–1208.
- Donner, A., and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold, London, U.K.
- Donner, A., Piaggio, G., Villar J., Pinol, A., Al-Mazrou, Y., Ba'aqeel, H., Bakketeg, L., Belizan, J. M., Berendes, H., Carroli, G., Farnot, U., and Lumbiganon, P. (1998). Methodological considerations in the design of the WHO Antenatal Care Randomized Controlled Trial. *Pediatric and Perinatal Epidemiology*, **12**, 59–74.
- Dornan, T. L., Heller, S. R., Peck, G. M., and Tattersol, R. B. (1991). Double-blind evaluation of efficacy and tolerability of Metformin in NIDDM. *Diabet. Care*, **14**, 342–344.
- Dougherty, T. B., Prosche, V., and Thall P. F. (1999). Maximal tolerable dose of nalnefene in patients receiving epidural fentanyl and dilute bupivacaine for postoperative analgesia. *Anesthesiology*, **92**, 1010–1016.
- Druker, B. J., Talpaz, M., Resta, D. J., et al. (2001). Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *New Engl. J. Med.*, **344**, 1031–1037.
- Dubey, S. D. (1991). Some thoughts on the one-sided and two-sided tests. *J. Biopharm. Stat.*, **1**, 139–150.
- Dunnett, C. W., (1995). A multiple comparison procedure for comparing several treatments with a control. *J. Am. Stat. Assoc.*, **50**, 1096–1121.
- Dunnett, C. W., and Gent, M. (1977). Significance testing to establish equivalence between treatments with special reference to data in the form of 2×2 table. *Biometrics*, **33**, 593–602
- Dunnett, C. W., and Gent, M. (1996). An alternative to the use of two-sided tests in clinical trials. *Stat. Med.*, **15**, 1729–1738.
- Dunnett, C. W., and Goldsmith, C. H. (1995). When and how to do multiple comparisons. In *Statistics in the Pharmaceutical Industry*, Ed. by Buncher, C. R., and Tsay, J. Y. Dekker, New York.
- Dunnett, C. W., and Tamhane, A. J. (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Stat. Med.*, **10**, 939–947.
- Dunnett, C. W., and Tamhane, A. J. (1992). Comparisons between a new drug and active drug and placebo controls in an efficacy clinical trial. *Stat. Med.*, **11**, 1057–1063.
- Durham, L. K., Longini, I. M., Halloran, M. E., Clemens, J. D., Nizam, A., and Rao, M. (1998). Estimation of vaccine efficacy in the presence of waning: application to cholera vaccine. *Am. J. Epidemiol.*, **147**, 948–959.
- Durrleman, S., and Simon (1990). Planning and Monitoring of equivalence studies. *Biometrics*, **46**, 329–336.

- Ebbeling, C. B., and Clarkson, P. M. (1989). Exercise-induced muscle damage and adaptation. *Sport Med.*, **7**, 207–234.
- Ebbutt, A. F., and Frith, L. (1998). Practical issue in equivalence trials. *Stat. Med.*, **17**, 1691–1701.
- Echt, D. S., Liebson, P. R., Mitchell, L. B., Peters, R. W., Obias-Manno, D., Barker, A. H., Arensberg, D., Baker, A., Friedman, L., Greene, H. L., Huther, M. L., Richardson, D. W., and the CAST Investigators (1991). Mortality and morbidity in patients receiving encainide and flecainide or placebo. *New Engl. J. Med.*, **324**, 781–788.
- Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, **58**, 403–417.
- Eichenwald, K., and Kolata, G. (1999). A doctor's trials turned into fraud. *New York Times*, May 17.
- Elkeles, R. S. (1991). The effects of oral hypoglycemic drug on serum lipids and lipoproteins in NIDDM. *Diabete Metab.*, **17**, 197–200.
- Ellenburg, J. H. (1990). Biostatistical collaboration in medical research. *Biometrics*, **46**, 1–32.
- Ellenberg, S. S. (2001). Safety considerations for new vaccine development. *Pharmacoepidemiology and Drug Safety*, **10**, 1–5.
- Ellenberg, S. S., and Dixon, D. O. (1994). Statistical issues in designing clinical trials of AIDS treatments and vaccines. *Journal of Statistical Planning and Inference*, **42**, 123–135.
- Ellenberg, S. S., and Temple, R. (2000). Placebo-controlled trials and active-controlled trials in the evaluation of new treatment, part 2: Practical issue and specific cases. *Ann. Int. Med.*, **133**, 464–470.
- Ensign, L. G., Gehan, E. A., Kamen, D. S., and Thall, P. F. (1994). An optimal three-stage design for phase II clinical trials. *Stat. Med.*, **13**, 1727–1736.
- Espland, M. A., Bush, T. L., Mebane-Sims, I., Stefanick, M. L., Johnson, S., Sherwin, R., and Waclawiw, M. (1995). Rationale, design, and conduct of the PEPI trial. *Controlled Clin. Trials*, **16**, 3S–19S.
- ESVEM (1989). The ESVEM trial: Electrophysiologic study versus electrocardiographic monitoring for selection of antiarrhythmic therapy of ventricular tachyarrhythmias. *Circul.*, **79**, 1354–1360.
- ESVEM (1993). Determinants of predicted efficacy of antiarrhythmic drugs in the Electrophysiologic study versus electrocardiographic monitoring trials. *Circul.*, **87**, 323–329.
- Evans, W. E., and McLeod, H. L. (2003). Pharmacogenomics—drug disposition, drug targets, and side effects. *New Engl. J. Med.*, **348**, 538–549.
- Fairweather, W. R. (1994). Statisticians, the FDA and a time of transition. Presented at Pharmaceutical Manufacturers Associated Education and Research Institute Training Course in Non-clinical Statistics, Georgetown University Conference Center, February 6–8, 1994, Washington, DC.
- Farlow, M., Gracon, S. I., Hershey, L. A., Lewis, K. W., Sadowsky, C. H., and Dolan-Ureno, J. (1992). A controlled trials of tacrine in Alzheimer's disease. *J. Am. Med. Assoc.*, **268**, 2523–2529.
- Farrington, C. P., and Manning, G. (1990). Test statistics and sample size formulae for comparing binomial trials with null hypothesis of no-zero risk difference of non-unity relative risk. *Stat. Med.*, **9**, 1447–1454.
- Fazzari, M., Heller, G., and Scher, H. I. (2000). The phase II/III toward the proof of efficacy in cancer clinical trials. *Controlled Clin. Trials*, **21**, 360–368.
- FDA (1997a). Guidance for Industry—*For the Evaluation of Combination Vaccines for Preventable Diseases: Production, Testing and Clinical Studies*. Center for Biologics Evaluation and Research, the United States Food and Drug Administration, Rockville, MD.
- FDA (1997b). Code of Federal Regulations, Title 21, Chapter 1, Part 11—*Electronic Records, Electronic Signatures*. The U.S. Food and Drug Administration, Rockville, MD.
- FDA (1988). *Guideline for the Format and Content of the Clinical and Statistical Sections of New Drug Applications*, U.S. Food and Drug Administration, Rockville, MD.

- FDA (1999). *Guidance for Industry—Computerized System Used in Clinical Trials*. The U.S. Food and Drug Administration, Rockville, MD.
- FDA (2001a). Guidance for Industry on *Statistical Approaches to Establishing Bioequivalence*. Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, MD.
- FDA (2001b). *Bioresearch Monitoring—Compliance Program Guidance Manual*. The U.S. Food and Drug Administration, Rockville, MD.
- FDA (2001c). *Draft Guidance on the Establishment and Operation of Clinical Trial Data Monitoring Committees*, Rockville, MD.
- FDA (2003a). Draft Guidance on *Developing Medical Imaging Drugs and Biological Products, Part 3: Design, Analysis and Interpretation of Clinical Studies*. The United States Food and Drug Administration, Rockville, MD.
- FDA (2003b). *Guidance for Industry on Part 11, Electronic Records; Electronic Signatures—Scope and Application*. The U.S. Food and drug Administration, Rockville, MD.
- Feigl, P., Blumestein, B., Thompson, I., Crowley, J., Wolf, M., Kramer, B. S., Coltman, C. A. Jr., Brawley, O. W., and Ford, L. G. (1995). Design for the Prostate Cancer Prevention Trial (PCPT). *Controlled Clin. Trial*, **6**, 150–163.
- Feingold, M., and Gillespie, B. W. (1996). Cross-over trials with censored data. *Stat. Med.*, **15**, 953–967.
- Feinstein, A. R. (1977). *Clinical Biostatistics*. C. V. Mosby, St. Louis, MO.
- Feinstein, A. R. (1989). Models, methods, and goals. *J. Clin. Epidemiol.*, **42**, 301–308.
- Feldt, L. S., and Mahmoud, M. W. (1958). Power function charts for specifying numbers of observations in analyses of variance of fixed effects. *Ann. Math. Stat.*, **29**, 871–877.
- Feuer, E. J., and Kessler, L. G. (1989). Test statistic and sample size for a two-sample McNemar test. *Biometrics*, **45**, 629–636.
- Fijal, B. A., Hall, J. M., and Witte, J. S. (2000). Clinical trials in the genomic era: effects of protective genotypes on sample size and duration of trial. *Controlled Clin. Trials*, **21**, 7–20.
- Fisher, L. D. (1990). *Biostatistics: Methodology for the Health Sciences*. Wiley, New York.
- Fisher, L. D. (1991). The use of one-sided tests in during trials: An FDA advisory committee member's perspective. *J. Biopharm. Stat.*, **1**, 151–156.
- Fisher, L. D., Gent, M., and Buller, H. R. (2001). Active-control trials: how would a new agent compare with placebo? A method illustrated with clopidogrel, aspirin, and placebo. *Am. Heart J.*, **141**, 26–32.
- Fisher, R. A. (1935). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.
- Fisher, R. A. (1947). *The Design of Experiments*, 4th Ed. Oliver and Boyd, Edinburgh.
- Fisher, R. A., and Mackenzie, W. A. (1923). Studies in crop variation II: The manurial response of different potato varieties. *J. Agri. Sci.*, **13**, 311–320.
- Fitzmaurice, G. M., Molenberghs, G., and Lipsitz, S. R. (1995). Regression models for longitudinal binary responses with informative drop-outs. *J. Roy. Stat. Soc.*, **57**, 691–704.
- Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*, 2nd Ed. Wiley, New York.
- Fleiss, J. L. (1986a). *The Design and Analysis of Clinical Experiments*. Wiley, New York.
- Fleiss, J. L. (1986b). Analysis of data from multiclinic trials. *Controlled Clin. Trials*, **7**, 267–275.
- Fleiss, J. (1987). Some thoughts on two-tailed tests (Letter to the Editor). *Controlled Clin. Trials*, **8**, 394.
- Fleming, T. R. (1982). One sample multiple testing procedure for phase II clinical trials. *Biometrics*, **38**, 143–151.
- Fleming, T. R. (2000). Design and interpretation of equivalence trials. *Am. Heart J.*, **139**, S172–S176.

- Fleming, T. R., O'Fallon, J. P., O'Brien, P. C., and Harrington, D. P. (1980). Modified Kolmogorov-Smirnov test procedures with application to arbitrary right-censored data. *Biometrics*, **36**, 607–626.
- Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *J. Psychiat. Rev.*, **12**, 189–198.
- Fong, D. Y. T. (2001). Data management and quality assurance. *Drug Info. J.*, **35**, 839–844.
- Foulds, G. A. (1958). Clinical research in psychiatry. *J. Mental Sci.*, **104**, 259–265.
- France, L. A., Lewis, J. A., and Kay, R. (1991). The analysis of failure time data in crossover studies. *Stat. Med.*, **10**, 1099–1113.
- Freedman, L., Anderson, G., Kipnis, V., Prentice, R., Wang, C. Y., Rossouw, J., Witten, J., and DeMets, D. (1996). Approaches to monitoring the results of long-term disease prevention trials: Examples from the Women's Health Initiative. *Controlled Clin. Trials*, **17**, 509–525.
- Freedman, L. S. (1982). Tables of the number of patients required in clinical trials using logrank test. *Stat. Med.*, **1**, 121–129.
- Freedman, L. S., and White, S. J. (1976). On the use of Pocock and Simon's method for balancing treatment numbers over prognostic factors in the controlled clinical trials. *Biometrics*, **32**, 691–694.
- Frey, S. E., Dagan, R., Ashur, Y., Chen, X., Ibarra, J., Kollaritsch, H., Mazur, M. H., Poland, G. A., Reisinger, K., Walter, E., Van Damme, P., Braconier, J. H., Uhnoo, I., Wahl, M., Blatter, M. M., Clements, D., Greenberg, D., Jacobson, R. M., Norrby, S. R., Rowe, M., Shouval, D., Simmons, S. S., van Hattum, J., Wennerholm, S., O'Brien Gress, J., Chan, I. S. F., and Kuter, B. (1999). Interference of antibody production on hepatitis B surface antigen in a combination hepatitis A/hepatitis B vaccine. *Journal of Infectious Diseases*, **180**, 2018–2022.
- Frey, S. E., Newman, F. K., Cruz, J., Shelton, B., Tennant, J. M., Polach, T., Rothman, A. L., Kennedy, J. S., Wolf, M., Belshe, R. B., and Ennis, F. A. (2002). Dose-related effects of smallpox vaccine. *New Engl. J. Med.*, **346**, 1275–1280.
- Frick, M. H., Elo, O., Haapa, K., Heinonen, O. P., Heinsalmi, P., Helo, P., Huttunen, J. K., Kaitaniemi, P., Koskinen, P., Manninen, V., Maenpaa, H., Malkonen, M., Manttari, M., Norola, S., Pasternack, A., Pikkarainen, J., Romo, M., Sjöblom, T., and Nikkila, E. A. (1987). Helsinki heart study: Primary prevention trial with gemfibrozil in middle-aged men with dyslipidemia. *New Engl. J. Med.*, **317**, 1237–1245.
- Friedman, L. M., Furberg, C. D., and DeMets, D. L. (1981). *Fundamentals of Clinical Trials*. 3rd Ed. Wiley, New York.
- Gail, M. H. (1985). Applicability of sample size calculations based on a comparison of proportions for use with the logrank test. *Controlled Clin. Trials*, **6**, 112–119.
- Gail, M. H., Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, **41**, 361–372.
- Gail, M. H., Wieand, S., and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitting covariates. *Biometrika*, **71**, 431–444.
- Gasparini, M., and Eisele, J. (2000). A curve-free method for phase I clinical trials. *Biometrics*, **56**, 609–615.
- Gehan, E. A. (1961). The determination of the number of patients required in a follow-up trial of a new chemotherapeutic agent. *J. Chron. Dis.*, **13**, 346–353.
- Gehan, E. A. (1965a). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, **52**, 203–223.
- Gehan, E. A. (1965b). A generalized two-sample Wilcoxon test for doubly censored data. *Biometrika*, **52**, 1965.
- Geller, N. L. (1994). Discussion of interim analysis: The alpha spending approach. *Stat. Med.*, **13**, 1353–1356.

- George, S. L., and Desu, M. M. (1973). Planning the size and duration of a clinical trial studying the time to some critical event. *J. Chron. Dis.*, **27**, 15–24.
- Gibson, D., Harvey, A. J., Everett, V., and Parmar, M. K. B. (1994). Is double entry necessary? The CHART trials. *Controlled Clinical Trials*, **15**, 482–488.
- Gilbert, G. S. (1992). *Drug Safety Assessment in Clinical Trials*. Dekker, New York.
- Giugliano, D., Quatraro, A., Consoli, G., Minei, A., Ceriello, A., De Rosa, N., and D’Onofrio, F. (1993). Metformin for obese, insulin-treated diabetic patients: Improvement in glycemic control and reduction of metabolic risk factors. *Eur. J. Clin. Pharmacol.*, **44**, 107–112.
- Glantz, S. A. (1987). *Primer of Biostatistics*, 2d Ed. McGraw-Hill, New York.
- Glanz, K., Fiel, S. B., Swartz, M. A., and Francis, M. E. (1984). Compliance with an experimental drug regimen for treatment of asthma: Its magnitude, importance, and correlates. *J. Chron. Dis.*, **37**, 815–824.
- Goldberg, J. D., and Koury, K. J. (1990). Design and Analysis of Multicenter Trials. In *Statistical Methodology in the Pharmaceutical Industry*, Ed. by Berry, D. Dekker, New York, 201–237.
- Goldstein, I., Lue, T. F., Padma-Nathan, H., Rosen, R. C., Steers, W. D., and Wicker, P. A. (1998). Oral sildenafil in the treatment of erectile dysfunction. *New Engl. J. Med.*, **338**, 1397–1404.
- Goodman, S. N. (1992). A comment on replication, p-values and evidence. *Stat. Med.*, **11**, 875–879.
- Goodman, S. N., Zahurak, M. L., and Piantadosi, S. (1995). Some practical improvements in the continual reassessment method for phase I studies. *Stat. Med.*, **14**, 1149–1161.
- Gormley, G. J., Stoner, E., Bruskewitz, R. C., Imperato-McGinley, J., Walsh, P. C., McConnell, J. D., Andriole, G. L., Geller, J., Bracken, B. R., Tenover, J. S., Vaughan, E. D., Pappas, F., Taylor, A., Binkowitz, B., Ng, J., for the Finasteride Study Group (1992). The effect of finasteride in men with benign prostatic hyperplasia. *New Engl. J. Med.*, **327**, 1185–1191.
- Gould, A. L. (1992). Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. *Stat. Med.*, **11**, 55–66.
- Gould, A. L. (1995a). Planning and revising the sample size for a trial. *Stat. Med.*, **14**, 1039–1051.
- Gould, A. L. (1995b). Group sequential extensions of standard bioequivalence testing procedure. *J. Pharmacokinetics Biopharmaceutics*, **23**, 57–85.
- Gould, A. L., and Shih, W. J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Comm. Stat. Theory Meth.*, **21**, 2833–2853.
- Grady, D. (2000). Breast cancer researchers admits falsifying data. *New York Times*.
- Graybill, F., and Wang, C. M. (1980). Confidence intervals on nonnegative linear combinations of variances. *J. Am. Stat. Assoc.*, **75**, 869–873.
- Green, S. J., and Dahlberg, S. (1992). Planned versus attained designs in phase II clinical trials. *Stat. Med.*, **11**, 853–862.
- Greenberg, R. P. and Fisher, S. (1994). Seeing through the double-masked design: a commentary. *Controlled Clin. Trials*, **15**, 244–246.
- Greene, J. G., and Hart, D. M. (1987). Evaluation of a psychological treatment programme for climacteric women. *Maturitas*, **9**, 41–48.
- Grier, M. T., and Meyers, D. G. (1993). So much writing, so little science, a review of 37 years of literature on edetate sodium chelation therapy. *Annals of Pharmacotherapy*, **27**, 1504–1509.
- Grobler, A. C., Harris, S. L., and Jooste, H.-L. (2001). The role of the statistician in the data management process. *Drug Info. J.*, **35**, 665–670.
- Grizzle, J. E., and Allen, D. M. (1969). Analysis of growth and dose response curves. *Biometrics*, **25**, 357–381.
- GUSTO I (1993). An investigational randomized trial comparing four thrombolytic strategies for acute myocardial function. The GUSTO investigators. *New Engl. J. Med.*, **329**, 673–682.

- GUSTO III Investigator (1997). A comparison of reteplase with alteplase for acute myocardial infarction. *New Engl. J. Med.*, **337**, 1118–1123.
- Guyatt, G., Sackett, M. D., Taylor, D. W., Chong, M. D., Roberts, M. S., and Pugsley, M. D. (1986). Determining optimal therapy-randomized trials in individual patients. *New Engl. J. Med.*, **314**, 889–892.
- Haaland, P. D. (1991). *Experimental Design in Biotechnology*. Dekker, New York.
- Halpern, S. D., Karlaawish, J. H. T., and Berlin, J. A. (2002). The continued unethical conduct of underpowered clinical trials. *J. Am. Med. Assoc.*, **288**, 358–362.
- Hand, D. J., and Crowder, M. J. (1996). *Practical Longitudinal Data Analysis*. Chapman and Hall, London.
- Harris, E. K., and Albert, A. (1991). *Survivorship Analysis for Clinical Studies*. Dekker, New York.
- Harville, D. A. (1977). Maximum likelihood approaches to variance components estimation and to related problems. *J. Am. Stat. Assoc.*, **72**, 320–338.
- Haseman, J. K. (1978). Exact sample sizes for use with the Fisher-Irwin test for 2×2 tables. *Biometrics*, **34**, 106–109.
- Hasselballd, V., and Kong, D. F. (2001). Statistical methods for comparison to placebo in active control trials. *Drug Info. J.*, **35**, 435–449.
- Havlir D. V., Marschner, I. C., Hirsch, M. S., Collier, A. C., Tebas, P., Bassett, R. L., Ioannidis, J. P. A., Holohan, M. K., Leavitt, R., Boone, G., Richman, D. D., for the AIDS. Clinical Trials Group Study 343 Team (1998). Maintenance antiretroviral therapies in HIV-infected subjects with undetectable plasma HIV RNA after triple-drug therapy. *New Engl. J. Med.*, **339**, 1261–1268.
- Haybittle, J. L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *Br. J. Radiol.*, **44**, 793–797.
- Heart Special Project Committee (1988). Organization, review and administration of cooperative studies (Greenberg report): A report from the Heart Special Project Committee to the National Advisory Council, May 1967. *Controlled Clin. Trials*, **9**, 137–148.
- Helms, R. W. (2001). Data quality issues in electronic data capture. *Drug Info. J.*, **35**, 827–837.
- Helms, R. W., Fitzmartin, R., Fillo, P., Miller, S., Talley, L., and Murphy, R. (2001). Metrics and best practices in clinical data management: conclusions of a DIA roundtable workshop. *Drug Info. J.*, **35**, 681–694.
- Henderson, J. D. (1993). Bioequivalence and bioavailability. Invited presentation at Generic Drug Approvals Workshop. Regulatory Affairs Professionals Society, May 4, 1993, Reston, VA.
- Hennekens, C. H., Buring, J. E., Manson, J. E., Stampfer, M., Rosner, B., Cook, N. R., Belanger, C., LaMotte F., Gaziano, J. M., Ridker, P. M., Willett, W., and Peto, R. (1996). Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease. *New Engl. J. Med.*, **334**, 1145–1149.
- Herrington, D. M., Reboussin, D. M., Broshinan, K. B., Sharp, P. C., Shumaker, S. A., Snyder, T. E., Furberg, C. D., Kowalchuk, G. J., Stuckey, T. D., Rogers, W. J., Givens, D. H., and Waters, D. (2000a). Effects of estrogen replacement on the progression of coronary artery atherosclerosis. *New Engl. J. Med.*, **343**, 522–529.
- Herrington, D. M., Reboussin, D. M., Klein, K. P., Sharp, P. C., Shumaker, S. A., Snyder, T. E., and Geisinger, K. R. (2000b). The Estrogen Replacement and Atherosclerosis (ERA) study: study design and baseline characteristic of the cohort. *Controlled Clin. Trials*, **21**, 257–285.
- Herrington, D. M., Howard, T. D., Hawkins, G. A., Reboussin, D. M., Xu, J., Zheng, S. L., Brosnan, K. B., Meyers, D. A., and Bleeker, E. R. (2002). Estrogen receptor polymorphism and effects of estrogen replacement of high-density lipoprotein cholesterol in women with coronary disease. *New Engl. J. Med.*, **346**, 967–974.

- Heyd, J. M., and Carlin, B. P. (1999). Adaptive design improvement in the continual reassessment method for phase I studies. *Stat. Med.*, **18**, 1307–1321.
- Hill, A. B. (1962). *Statistical Methods in Clinical and Preventive Medicine*. Oxford University Press, New York.
- Hines, L. K., Laird, N. M., Hewitt, P., and Chalmers, T. C. (1989). Meta-analysis of empirical long-term antiarrhythmic therapy after myocardial infarction. *J. Am. Med. Assoc.*, **262**, 3037–3040.
- Hochberg, Y. (1988). A sharper Bonferroni's procedure for multiple tests of significance. *Biometrika*, **75**, 800–803.
- Hochberg, Y., and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. Wiley, New York.
- Hollander, M., and Wolfe, D. A. (1973). *Nonparametric Statistical Methods*. Wiley, New York.
- Holm, S. (1979). A sharper Bonferroni procedure for multiple tests of significance. *Scand. J. Stat.*, **6**, 65–70.
- Holt, J. D., and Prentice, R. L. (1974). Survival analysis in twin studies and matched-pair experiments. *Biometrika*, **61**, 17–30.
- Hoover, D. R. (1996). Extension of life table to repeating and changing events. *Am. J. Epidemiol.*, **143**, 1266–1276.
- Hortobagyi, G. N. (1998) Treatment of breast cancer. *New Engl. J. Med.*, **329**, 974–984.
- Horton, R. (2000). Retraction: Interferon alfa-2b . . . in Bechet's disease. *Lancet*, **356**, 1292.
- Horton, R. (2001). The clinical trial: Deceitful, disputable, unbelievable, unhelpful, and shameful—what next. *Controlled Clin. Trials*, **22**, 593–604.
- Hosmer, D. W., and Lemeshow, S. (2000). *Applied Logistic Regression*. 2nd Ed. Wiley, New York.
- Howe, W. G. (1974). Approximate confidence limits on the mean of X + Y where X and Y are two tabled independent random variables. *J. Am. Stat. Assoc.*, **69**, 789–794.
- Hsu, J. C., Hwang, J. T. G., Liu, H. K., and Ruberg, S. J. (1994). Confidence intervals associated with tests for bioequivalence. *Biometrika*, **81**, 103–114.
- Hsu, J. P. (1983). The assessment of statistical evidence from active control clinical trials. *Proc. of Biopharmaceutical Section of the American Statistical Association*, 12–17.
- Hsuan, A. T. M. and Genyn, P. (2003). Global database and system. In *Encyclopedia of Biopharmaceutical Statistics*. 2nd Ed. Ed. by Chow, S.C. Dekker, New York.
- Hsueh, H. M., Liu, J. P., and Chen, J. J. (2001). Unconditional exact tests for equivalence or noninferiority for paired binary endpoints. *Biometrics*, **57**, 478–483.
- Hsueh, H. M., Yao, T. J., and Liu, J. P. (2002). Evaluation of equivalence and non-inferiority tests in the comparison of two survival functions. An invited presentation at the International Chinese Statistical Association (ICSA) 2002 Applied Physics Symposium, June 6–8, 2002, Plymouth Meeting, PA.
- Hughes, M. D. (1993). Stopping guidelines for clinical trials with multiple treatments. *Stat. Med.*, **12**, 901–913.
- Hughes, M. D. (1997). Power considerations for clinical trials using multivariate time-to-event data. *Stat. Med.*, **16**, 865–882.
- Hughes, M. D., and Pocock, S. J. (1988). Stopping rules and estimation problems in the clinical trials. *Stat. Med.*, **7**, 1231–1242.
- Hui, S. L. (1984). Curve fitting for repeated measurements made at irregular time points. *Biometrics*, **40**, 691–697.
- Huitson, A., Poloniecki, J., Hews, R., and Barker, N. (1982). A review of crossover trials. *Statistician*, **31**, 71–80.
- Hung, J. H. M. (1992). On identifying a positive dose-response surface for combination agents. *Stat. Med.*, **11**, 703–711.

- Hung, J. H. M. (1994). Correspondence: Test for the existence of a desirable dose combination. *Biometrics*, **50**, 307–308.
- Hung, J. H. M. (1996). Global tests for combination drug studies in factorial trials. *Stat. Med.*, **15**, 233–247.
- Hung, H. M. J. (2001). Noninferiority: A dangerous toy. *ICSA Bulletin*, January, 27–29.
- Hung, J. H. M., Chi, G. Y. H., and Lipicky, R. J. (1993). Testing for the existence of a desirable dose combination. *Biometrics*, **49**, 85–94.
- Hung, J. H. M., Ng, T. H., Chi, G. Y. H., and Lipicky, R. J. (1990). Response surface and factorial designs for combination antihypertensive drugs. *Drug Info. J.*, **24**, 371–378.
- Hung, J. H. M., Ng, T. H., Chi, G. Y. H., and Lipicky, R. J. (1989). Testing for the existence of a dose combination beating its components. *Proc. of Biopharmaceutical Section of the American Statistical Association*, Alexandria, VA, 53–59.
- Hung, J. H. M., Chi, G. Y. H., and O'Neill, R. T. (1995). Efficacy evaluation for monotherapies in two-by-two factorial trials. *Biometrics*, **51**, 1483–1493.
- Hung, H. M. J., O'Neill, R. T., Bauer, P., and Kohne, K. (1997). The behavior of the p-value when alternative hypothesis is true. *Biometrics*, **53**, 11–22.
- Huque, M. F., Dubey, S., and Fredd, S. (1989). Establishing therapeutic equivalence with clinical endpoints. *Proc. of Biopharmaceutical Section of the American Statistical Association*, Alexandria, VA, 46–52.
- Huque, M. F., and Dubey, S. (1990). Design and analysis for therapeutic equivalence clinic trials with binary clinic endpoints. *Proc. of Biopharmaceutical Section of the American Statistical Association*, Alexandria, VA, 91–97.
- Hurwitz, N. (1969). Admission to hospital due to drugs. *Br. Med. J.*, **1**, 536–539.
- Huster, W. J., and Louv, W. C. (1992). Demonstration of the reproducibility of treatment efficacy from a single multicenter trial. *J. Biopharmaceut. Stat.*, **2**, 219–238.
- Huster, W. J., Brookmeyer, R., and Self, S. G. (1989). Model paired survival data with covariates. *Biometrics*, **45**, 145–156.
- Huwiler-Münsterer, K., Jüni, P., and Junker, C. (2002). Quality of reporting of randomized trials as a measure of methodologic quality. *Journal of the American Medical Association*, **287**, 2801–2804.
- Hwang, I. K. (2001). Noninferiority trials, dose sloppiness bias towards no difference. *ICSA Bulletin*, January, 25–27.
- Hyslop, T., Hsuan, F., and Holder, D. J. (2000). A small sample confidence interval approach to assess individual bioequivalence. *Stat. Med.*, **19**, 2885–2897.
- ICH (1995). International Conference on Harmonization Tripartite Guideline E2A *Clinical Safety Data Management: Definitions and Standards for Expedited Reporting*.
- ICH (1994). International Conference on Harmonization Tripartite Guideline E4 *Guideline on Dose-response Information to Support Drug Registration*.
- ICH (1996). International Conference on Harmonization Tripartite Guideline E3 *Structure and Content of Clinical Study Reports*.
- ICH (1996). International Conference on Harmonization Tripartite Guideline E6 *Good Clinical Practice: Consolidated Guidance*.
- ICH (1997). International Conference on Harmonization Tripartite Guideline E8 *General Principles for Clinical Trials*.
- ICH (1998). International Conference on Harmonization Tripartite Guideline E5 *Ethnic Factors in the Acceptability of Foreign Data*; The U.S. Federal Register, **83**, 31790–31796.
- ICH (1998). International Conference on Harmonisation. Guideline E9 *on Statistical Principles for Clinical Trials*.

- ICH (1999). International Conference on Harmonisation. Guideline E10 on *Choice of Control Group in Clinical Trials*.
- IDSA (1990). Infectious Diseases Society of America: Guidelines for the use of antimicrobial agents in neutropenic patients with unexplained fever. *J. Infect. Dis.*, **161**, 381–396.
- IHS (1990). The design, analysis, and reporting of clinical trials on the empirical antibiotic management of the neutropenic patient: Report of a consensus panel for the Immunocompromised Host Society. *J. Infect. Dis.*, **161**, 397–401.
- Ioannidis, J. P. A., Haidich, A. B., Pappa, M., Kokori, S. I., Tektonidou, M. G., Contopoulos-Ioannidis, D. G., and Lau, J. (2001). Comparison of evidence of treatments effects in randomized and non-randomized studies. *J. Am. Med. Assoc.*, **286**, 821–830.
- Iman, R. L., Quade, D., and Alexander, D. A. (1975). Exact probability levels for the Kruskal–Wallis test, *Selected Tables in Mathematical Statistics*, **3**, 329–384.
- Ishizuka, N., and Ohashi, Y. (2001). The continual reassessment method and its applications: A Bayesian methodology for phase I cancer clinical trials. *Stat. Med.*, **20**, 2661–2681.
- ISIS-2 Group (1988). Randomized trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction. *Lancet*, **13**, 349–360.
- Jadad, A. R., Moore, R. A., Carroll, D., Jenkinson, C., Reynolds, D. J., Gavahan, D. J., and McQuay, H. J. (1996). Assessing the quality of reports of randomized clinical trial: is blinding necessary. *Controlled Clin. Trials*, **17**, 1–12.
- Jennison, C., and Turnbull, B. (1989). Interim analysis: the repeated confidence interval approach (with discussion). *J. Roy. Stat. Soc., B* **51**, 305–361.
- Jennison, C., and Turnbull, B. W. (1990). Statistical approaches to interim monitoring of medical trials: A review and commentary. *Stat. Sci.*, **5**, 299–317.
- Jennison, C., and Turnbull, B. (1991). Exact calculation for sequential, *t*, chi-square, and *F* tests. *Biometrika*, **78**, 133–141.
- Jennison, C., and Turnbull, B. (1993). Sequential equivalence testing and repeated confidence intervals, with application to normal and binary responses. *Biometrics*, 31–44.
- Jennison, C., and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC, New York.
- John, P. W. M. (1971). *Statistical Design and Analysis of Experiments*. Macmillan, New York.
- Jones, B., Jarvis, P., Lewis, J. A., and Ebbutt, A. F. (1996). Trials to assess equivalence: the importance of rigorous methods. *Br. Med. J.*, **313**, 36–39.
- Jones, B., and Kenward, M. G. (1989). *Design and Analysis of Crossover Trials*. Chapman-Hall, London.
- Ju, H. L. (2002). Good data management practices. Presented at National Taipei Medical University, Taipei, Taiwan, August.
- Jula, A., Marniemi, J., Huupponen, R., Virtanen, A., Rasras, M., and Ronnemaa, T. (2002). Effects of diet and simvastatin on serum lipids, insulin, and antioxidants in hypercholesterolemic men. *J. Am. Med. Assoc.*, **287**, 598–605.
- Jung, S. H., and Su, J. Q. (1995). Non-parametric estimation for the difference of ratio of median failure times for paired observations. *Stat. Med.*, **14**, 275–281.
- Jüni, P., Altman, D. G., and Egger, M. (2001). Assessment of the quality of randomized controlled trials, Chapter 5 in *Systematic Reviews in Health Care: Meta-analysis in Context*, Ed. by Egger, M., Smith, G. D., and Altman, D. G., British Medical Journal Publishing Group, London, U.K.
- Kalbfleisch, J. D., and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.
- Kalbfleisch, J. D., and Street, J. O. (1990). Survival analysis. In *Statistical Methodology in Pharmaceutical Science*, Ed. by Berry, D. A. Dekker, New York.

- Kallen, A., and Larson, P. (2001). Equivalence studies can not be used to claim equivalence of two active treatments. *ICSA Bulletin*, January, 33–35.
- Kang, S. H., and Chen, J. J. (2000). An approximate exact test for non-inferiority between two proportions. *Stat. Med.*, **19**, 2089–2100.
- Kantarjian, H. et al. (2002). Hematologic and Cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia. *New Engl. J. Med.*, **346**, 645–652.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Stat. Assoc.*, **53**, 457–481.
- Karlowski, T. R., Chalmers, T. C., Frenkel, L. D., Kapikian, A. Z., Lewis, T. L., and Lynch, J. M. (1975). Ascorbic acid for the common cold: A prophylactic and therapeutic trial. *J. Am. Med. Assoc.*, **231**, 1038–1042.
- Kastenbaum, M. A., Hoel, D. G., and Bowman, K. O. (1970). Sample size requirements: one way analysis of variance. *Biometrika*, **57**, 421–430.
- Kelly, P. J., and Lim, L. L. Y. (2000). Survival analysis for recurrent event data: An application to children infection diseases. *Stat. Med.*, **19**, 12–33.
- Kenward, M. G. (1987). A method for comparing profiles of repeated measurements. *Appl. Stat.*, **36**, 296–308.
- Kershner, R. P., and Federer, W. T. (1981). Two-treatment crossover design for estimating a variety of effects. *J. Am. Stat. Assoc.*, **76**, 612–618.
- Kessler, D. A. (1989). The regulation of investigational drugs. *New Engl. J. Med.*, **320**, 281–288.
- Kessler, D. A., and Feiden, K. L. (1995). Faster evaluation of vital drugs. *Sci. Am.*, **272**, 48–54.
- Keuls, M. (1952). The use of the studenized range in connection with an analysis of variance. *Euphytica*, **1**, 112–122.
- Khatri, C. G., and Patel, H. I. (1992). Analysis of a multicenter trial using a multivariate approach to a mixed linear model. *Comm. Stat. Theory Meth.*, **21**, 21–39.
- Kim, K. (1989). Point estimation following group sequential tests. *Biometrics*, **45**, 613–617.
- Kim, K., and DeMets, D. L. (1987). Confidence intervals following group sequential tests in clinical trials. *Biometrics*, **43**, 857–864.
- King, D. W., and Lashley, R. (2000). A quantifiable alternative to double data entry. *Controlled Clin. Trials*, **21**, 94–102.
- Kirshner, B. (1981). Methodological standards for assessing therapeutic equivalence, *J. Clin. Epidemiol.*, **44**, 839–849.
- Kleinbaum, D. G. (1996). *Survival Analysis, A Self-learning Text*. Springer-Verlag, New York.
- Knapp, M. J., Knopman, D. S., Solomon, P. R., Pendlebury, W. W., Davis, C. S., and Gracon, S. I. (1994). A 30-week randomized controlled trial of high-dose tacrine in patients with Alzheimer's disease. *J. Am. Med. Assoc.*, **271**, 985–991.
- Knudtson, M. L., Wyse, D. G., Galbraith, P. D., Brant, R., Hildebrand, K., Paterson, D., Richardson, D., Burkart, C., and Burgess, E., for PATCH Investigators (2002). Chelation therapy for ischemic heart disease. *J. Am. Med. Assoc.*, **287**, 481–486.
- Koch, G. C. (1991). One-sided and two-sided tests and p -values. *J. Biopharm. Stat.*, **1**, 161–170.
- Koch, G. C., and Bhapkar, V. P. (1982). Chi-square test. *Encyclopedia of Statistical Sciences*, Vol. 1, Ed. by Johnson, N. L. and Kotz, S. Wiley, New York.
- Koch, G. G., and Edwards, S. (1988). Clinical efficacy trials with categorical data. In *Biopharmaceutical Statistics for Drug Development*, Ed. by Peace, K. Dekker, New York.
- Koch, G. G., Carr, G. J., Amara, I. A., Stokes, M. E., and Uryniak, T. J. (1990). Categorical data analysis. In *Statistical Methodology in the Pharmaceutical Sciences*, Ed. by Berry, D. A. Dekker, New York.
- Koch, G. G., Amara, I. A., Davis, G. W., and Gillings, D. B. (1982). A review of some statistical methods for covariance analysis of categorical data. *Biometrics*, **38**, 563–595.

- Korn, E. L., and Simon, R. (1996). Data monitoring committees and problems of lower-than-expected accrual or event rates. *Controlled Clin. Trials*, **17**, 527–536.
- Korn, E. L., Midthune, D., Chen, T. T., Rubinstein, L. V., Christian, M. C., and Simon, R. (1999). Commentary. *Stat. Med.*, **18**, 2691–2692.
- Kramar, A., Lebecq A., and Candalh, E. (1999). Continual reassessment methods in phase I trials of the combination of two drugs in oncology. *Stat. Med.*, **18**, 1849–1864.
- Lachenburgh, P. A., and Lynch, C. J. (1998). Assessing screening tests: Extension of McNemar's test. *Stat. Med.*, **17**, 2207–2217.
- Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clin. Trials*, **2**, 93–113.
- Lachin, J. M. (1988a). Statistical properties of randomization in clinical trials. *Controlled Clin. Trials*, **9**, 289–311.
- Lachin, J. M. (1988b). Properties of simple randomization in clinical trials. *Controlled Clin. Trials*, **9**, 312–326.
- Lachin, J. M. (2000). Statistical considerations in the intent-to-treat principle, *Controlled Clin. Trials*, **21**, 167–189.
- Lachin, J. M., and Foulkes, M. A. (1986). Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to followup, noncompliance, and stratification. *Biometrics*, **42**, 507–519.
- Lachin, J. M., Matts, J. P., and Wei, L. J. (1988). Randomization in clinical trials: Conclusions and recommendations. *Controlled Clin. Trials*, **9**, 365–374.
- Laird, N. M. (1988). Missing data in longitudinal studies. *Stat. Med.*, **7**, 305–315.
- Lamborn, K. R. (1983). Some practical issues and concerns in active control clinical trials. *Proc. of Biopharmaceutical Section of the American Statistical Association*, Alexandria, VA, 8–12.
- Lakatos, E. (1986). Sample size determination in clinical trials with time-dependent rates of losses and noncompliance. *Controlled Clin. Trials*, **7**, 189–199.
- Lakatos, E. (1988). Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrika*, **44**, 229–241.
- Lan, K. K. G., and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, **70**, 659–663.
- Lan, K. K. G., and DeMets, D. L. (1989). Group sequential procedures: Calendar versus information time. *Stat. Med.*, **8**, 1191–1198.
- Lan, K. K. G., Simon, R., and Halperin, M. (1982). Stochastically curtailed testing in long-term clinical trials. *Comm. Stat.*, **C1**, 207–219.
- Lan, K. K. G., and Wittes, J. (1988). The *B*-value: A tool for monitoring data. *Biometrics*, **44**, 579–585.
- Lan, K. K. G., and Zucker, D. M. (1993). Sequential monitoring of clinical trials: The role of information and Brownian motion. *Stat. Med.*, **12**, 753–765.
- Lasagna, L. (1975). *Combination Drugs: Their Use and Regulations*. Stratton Intercontinental Medical Book, New York.
- Lasagna, L., Laties, V. G., and Dohan, J. L. (1958). Further studies on the "Pharmacology of placebo administration." *J. Clin. Invest.*, **37**, 533–537.
- Laska, E. M., and Meisner, M. (1985). A variational approach to optimal two-treatment crossover designs: Applications to carryover effects methods. *J. Am. Stat. Assoc.*, **80**, 704–710.
- Laska, E. M., and Meisner, M. (1989). Testing whether an identified treatment is best. *Biometrics*, **45**, 1139–1151.
- Laska, E. M., and Meisner, M. (1990). Hypothesis testing for combination treatments. In *Statistical Issues in Drug Research and Development*, Ed. by Peace, K. L. Dekker, New York, 276–284.

- Laska, E. M., Meisner, M., and Kushner, H. B. (1983). Optimal crossover designs in the presence of carryover effects. *Biometrics*, **39**, 1089–1091.
- Lau, G. S. N. (2000). Monitoring and data quality assurance. Presented at the DIA/University of Hong Kong meeting, Hong Kong, November.
- Laubscher, N. F. (1960). Normalizing the concentral t and F distributions. *Ann. Math. Stat.*, **31**, 1105–1112.
- Lavori, P. W., Krause-Steinrauf, H., Brophy, M., Buxbaum, J., Crokroft, J., Cox, D. R., et al. (2002). Principles, organization, and operation of a DNA bank for clinical trials, a Department of Veteran Affairs cooperative study. *Controlled Clin. Trials*, **23**, 222–239.
- Lawless, J. F., Nadeau, C., and Cook, R. J. (1997). Analysis of mean and rate functions for recurrent events. In *Proceedings of the 1st Seattle Symposium on Biostatistics: Survival Analysis*, (Ed. Lin, D.Y., and Fleming, T.R.) Springer, New York.
- Leber, P. D. (1989). Hazards of Inference: The active control interpretation. *Epilepsia*, **30**, S57–S63.
- Lee, E. T., Desu, M. M., and Gehan, E. A. (1975). A Monte Carlo study of the power of some two-sample tests. *Biometrika*, **62**, 423–425.
- Lee, K. L., Califf, R. M., Simers, J., Werf, F. V., and Topol, E. J. (1994). Holding GUSTO up to the light. *Ann. Internal Med.*, **120**, 876–881.
- Lee, Y., Shao, J., and Chow, S. C. (2003). The modified large sample confidence interval for linear combinations of variance components: extension, theory, and application, under revision for the *Journal American Statistical Association*.
- Lee, Y., Shao, J., Chow, S. C., and Wang, H. (2002). Tests for inter-subject and total variabilities under crossover designs. *J. Biopharm. Stat.*, **12**, 503–534.
- Legedza, A. T. R., and Ibrahim, J. G. (2000). Longitudinal design for phase I clinical trials using the continual reassessment method. *Controlled Clin. Trials*, **21**, 574–588.
- Lee, E. T., and Wang, J. W. (2003). *Statistical Methods for Survival Analysis*, 3rd ed, Wiley, New York.
- Lehmann, E. L. (1953). The power of rank tests. *Annals of Math. Stat.*, **24**, 23–43.
- Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco, 132–334.
- Lemeshow, S., Hosmer, D. W., and Stewart, J. P. (1981). A comparison of sample size determination methods in the two-group trials where the underlying disease is rare. *Comm. Stat. Simu. Computa.*, **B10**, 437–449.
- Lepor, H., Williford, W. O., Barry, M. J., Brawer, M. K., Dixon, C. M., Gormley, G., Haakenson, C., Machi, M., Narayan, P., and Padley, R. J. (1996). The efficacy of terazosin, finasteride, or both in benign prostatic hyperplasia. *New Engl. J. Med.*, **335**, 533–539.
- Levine, R. J. (1987). The apparent incompatibility between informed consent and placebo-controlled clinical trials. *Clin. Pharmacol. Ther.*, **42**, 247–249.
- Levine, J. G. (1996). Analysis and presentation of clinical trial adverse events data. Presented at the Biopharmaceutical Section Workshop on Adverse Events, October 28–29, 1996, Bethesda, MD.
- Levine, J. G., and Szarfman, A. (1996). Standardized data structures and visualization tool: a way to accelerate the regulatory review of the integrated summary of safety of new drug applications. *Biopharmaceut. Rep.*, **4**, 12–17.
- Lewis, J. A. (1995). Statistical issues in the regulation of medicine. *Stat. Med.*, **14**, 127–136.
- Lewis, J. A., Jones, D. R., and Röhmel, J. (1995). Biostatistical methodology in clinical—A European guideline. *Stat. Med.*, **14**, 1655–1657.
- Liang, K. Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
- Lin, D. Y., and Ying, Z. L. (2001). Nonparametric tests for the gap time distributions of serial events based on censored data. *Biometrics*, **57**, 369–375.

- Lin, D. Y., Wei, L. J., Yang, I., and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *J. Royal Stat. Soc., Series B*, **62**, 711–730.
- Linde, K., Ramirez, G., Mulrow, C. D., Pauls, A., Weldenhammer, W., and Melchart, D. (1996). St John's wort for depression: an overview and meta-analysis of randomized clinical trials. *Br. Med. J.*, **313**, 253–258.
- Lindsey, J. K. (1993). *Models for Repeated Measurements*. Oxford Science, New York.
- Lipid Research Clinics Program (1984). The lipid research clinics coronary primary prevention trials results. I: Reduction in incidence of coronary heart disease. *J. Am. Med. Assoc.*, **251**, 351–364.
- Lipsitz, S. R., Fitzmaurice, G. M., Orav, E. J., and Laird, N. M. (1944). Performance of generalized estimating equations in practical situations. *Biometrics*, **50**, 270–279.
- Lisook, A. B. (1990). Audits of clinical studies: Policy and procedures. *J. Clin. Pharmacol.*, **30**, 296–302.
- Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated measures studies. *J. Am. Stat. Assoc.*, **90**, 1112–1121.
- Little, R. J. A., and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.
- Liu, G., and Liang, K. Y. (1995). Sample size calculations for studies with correlated observations. Unpublished manuscript.
- Liu, J. P. (1994). Invited discussion of “Individual bioequivalence: A problem for switchability” by S. Anderson. *Biopharmaceut. Rep.*, **2**, 7–9.
- Liu, J. P. (1995a). Letter to the editor on “Sample size for therapeutic equivalence based on confidence interval” by S. C. Lin, *Drug Info. J.*, **29**, 45–50.
- Liu, J. P. (1995b). Use of the replicated crossover designs in assessing bioequivalence. *Stat. Med.*, **14**, 1067–1078.
- Liu, J. P., and Chow, S. C. (1992). Sample size determination for the two one-sided tests procedure in bioequivalence. *J. Pharmacokinetics & Biopharmaceutics*, **20**, 101–104.
- Liu, J. P., and Chow, S. C. (1993). Assessment of bioequivalence for drugs with negligible plasma levels. *Biomet. J.*, **36**, 109–123.
- Liu, J. P., and Chow, S. C. (1995). Replicated crossover designs in bioavailability and bioequivalence studies. *Drug Info. J.*, **29**, 871–884.
- Liu, J. P., and Chow, S. C. (2002). Bridging studies in clinical development, *J. Biopharm. Stat.*, Vol. 12, 357–369.
- Liu, J. P., Hsueh, H-M., Hsieh, E., and Chen, J. J. (2002). Tests for equivalence or non-inferiority for paired binary data. *Stat. Med.*, **21**, 231–245.
- Liu, J. P., Hsueh, H-M., and Chen, J. J. (2002). Sample size requirement for evaluation of bridging evidence. *Biomet. J.*, **44**, 969–981.
- Liu, J. P., Hsueh, H-M., and Hsiao, C. F. (2002). Bayesian approach to evaluation of the bridging studies. *J. Biopharm. Stat.*, **12**, 401–408.
- Liu, J. P., and Weng, C. S. (1995). Bias of two one-sided tests procedures in assessment of bioequivalence. *Stat. Med.*, **14**, 853–862.
- Liu, K. J. (2002). A flexible design for multiple armed screening trials. Letter to the Editor. *Stat. Med.*, **21**, 625–627.
- Liu, P. Y., Dahlberg, S., and Crowley, J. (1993). Selection designs for pilot studies based on survival. *Biometrics*, **49**, 391–393.
- Liu, P. Y., Leblanc, M., and Desai, M. (1999). False positive rates of randomized phase II designs. *Controlled Clin. Trials*, **20**, 343–352.
- Lu, Y., and Bean, J. A. (1995). On the sample size for one-sided equivalence of sensitivities based on McNemar's test. *Stat. Med.*, **14**, 1831–1839.
- Lui, K. J., and Cumberland, W. G. (2001). Sample size determination for equivalence test using rate ratio of sensitivity and specificity in paired sample data. *Controlled Clin. Trials*, **22**, 373–389.

- Lundh, L. G. (1987). Placebo, relief, and health: A cognitive-emotional model. *Scand. J. Psycho.*, **28**, 128–143.
- Lyden, P., Brott, T., Tilley, B., Welch, K. M. A., Mascha, E. J., Levine, S., Haley, E. C., Grotta, J., Marler, J., and the NINDS TPA Stroke Study Group (1994). Improved reliability of the NIH stroke scale using video training. *Stroke*, **25**, 2220–2226.
- Madison, T., and Plaunt, M. (2003). Clinical data management. In *Encyclopedia of Biopharmaceutical Statistics*. 2nd Ed. Ed. by Chow, S. C. Dekker, New York, 181–188.
- Mahe, C., and Chevret, S. (2001). Analysis of recurrent failure times data: Should the baseline hazard be stratified. *Stat. Med.*, **20**, 3807–3815.
- Mainland, D. (1952). *Elementary Medical Statistics: The principles of quantitative medicine*, W. B. Saunders, Philadelphia.
- Makuch, R., and Johnson, M. (1989). Issues in planning and interpreting active control equivalence studies. *J. Clin. Epidemiol.*, **42**, 503–511.
- Makuch, R. W., and Johnson, M. (1990). Active control equivalence studies: Planning and interpretation. In *Statistical Issues in Drug Research and Development*, Ed. by Peace, K. E. Dekker, New York, 238–246.
- Makuch, R. W., and Simon, R. (1978). Sample size requirements for evaluating a conservative therapy. *Cancer Treat. Rep.*, **6**, 1037–1040.
- Manninen, V., Elo, O., and Frick, M. H. (1988). Lipid alterations and decline in the incidence of coronary heart disease in the Helsinki Heart Study. *J. Am. Med. Assoc.*, **260**, 641–651.
- Mantel, N. (1963). Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *J. Am. Stat. Assoc.*, **58**, 690–700.
- Mantel, N. (1967). Ranking procedure for arbitrarily restricted observations. *Biometrics*, **23**, 65–78.
- Mantel, N., and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst.*, **22**, 719–748.
- Margolies, M. E. (1994). Regulations of combination products. *Appl. Clin. Trials*, **3**, 50–65.
- Mariani, L., and Marubini, E. (1996). Design and analysis of phase II cancer clinical trials: A review of statistical methods and guidelines for medical researchers. *International Statistical Review*, **64**, 61–68.
- Marubini, E., and Valsecchi, M. G. (1995). *Analyzing Survival Data from Clinical Trials and Observational Studies*. Wiley, New York.
- Mason, J. W., for the ESVEM Investigators (1993a). A comparison of seven antiarrhythmic drugs in patients with ventricular tachyarrhythmias. *New Engl. J. Med.*, **329**, 452–458.
- Mason, J. W., for the ESVEM Investigators (1993b). A comparison of electrophysiologic testing versus Holter monitoring to predict antiarrhythmic-drug efficacy for ventricular tachyarrhythmias. *New Engl. J. Med.*, **329**, 445–451.
- Matts, J. P., and Lachin, J. M. (1988). Properties of permuted-block randomization in clinical trials. *Controlled Clin. Trials*, **9**, 327–344.
- McCullagh, P., and Nelder, J. A. (1983). Quasi-likelihood functions. *Ann. Stat.*, **11**, 59–67.
- McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd Ed. Chapman and Hall, New York.
- McFadden, E. (1998). *Management of Data in Clinical Trials*. Wiley, New York.
- McGuire, W. P., Hoskins, W. J., Brady, M. F., Kucera, P. R., Patridge, E. E., Look, K. Y., Clarke-Pearson, D. L., and Davidson, M. (1996). Cyclophosphamide and cisplatin compared with paclitaxol and cisplatin in patients with stage III and stage IV ovarian cancer. *New Engl. J. Med.*, **334**, 1–6.
- McHugh, R., and Matts, J. (1983). Post-stratification in the randomized clinical trial. *Biometrics*, **39**, 217–225.
- Medical Dictionary for Regulatory Activities (MedDRA®) Introductory Guide* (End User Manual) MedDRA Version 5.1 (2002), MedDRA Maintenance and Support Services Organization (MSSO).

- Medical Research Council (1931). Clinical trials of new remedies (annotations); *Lancet*, **2**, 304.
- Medical Research Council (1948). Streptomycin treatment of pulmonary: A Medical Research Council investigation. *Br Med J.*, **2**, 769–782.
- Meier, P. (1989). The biggest public health experiment ever, the 1954 field trial of the Salk poliomyelitis vaccine. In *Statistics: A Guide to the Unknown*, Ed. by Tanur, J. M., Mosteller, F., and Kruskal, W. H., 3rd Ed. Wadsworth, Belmont, CA, 3–14.
- Meinert, C. L. (1986). *Clinical Trials: Design, Conduct, and Analysis*. Oxford University Press, New York.
- Metzler, C. M. (1974). Bioavailability: A problem in equivalence. *Biometrics*, **30**, 309–317.
- Mezey, K. C. (1980). *Fixed Drug Combination—Rationale and Limitations*. International Congress and Symposium Series, No. 22. Royal Society of Medicine, London; Academic Press, London.
- Miao, L. L. (1977). Gastric freezing: An example of the evaluation of medical therapy for randomized trials. In *The Costs, Risks, and Benefits of Surgery*, Ed. by Bunker, J. P., Barnes, B. A., Mosteller, F. Oxford University Press, New York.
- Miller, A. B., Hogestraeten, B., Staquet, M., and Winkler, A. (1981). Reporting results of cancer treatment. *Cancer*, **47**, 207–214.
- Mielke, P., and McHugh, R. B. (1965). Two-way analysis of variance for the mixed model with disproportionate sub-class frequencies. *Biometrics*, **21**, 308–323.
- Mike, V., and Stanley, K. E. (1982). *Statistics in Medical Research: Methods and Issues, with Applications in Cancer Research*. Wiley, New York.
- Miller, R. G. Jr. (1981). *Survival Analysis*. Wiley, New York.
- Miller, M. E., Davis, C. S., and Landis, J. R. (1993). The analysis of longitudinal polytomous data: GEE and connections with weighted least squares. *Biometrics*, **49**, 1033–1044.
- Miller, R. G., Efron, B., and Brown, B. W. (1980). *Biostatistics Case Book*. Wiley, New York.
- Møller, S. (1995). An extension of the continual reassessment methods using a preliminary up-and-down design in a dose finding study in cancer patients, in order to investigate a greater range of doses. *Stat. Med.*, **14**, 911–922.
- Moher, D., Jones, A., and Lepage, L. (2001). Use of the CONSORT statement: and quality of reports of randomized trials: A comparative before-and-after evaluation. *J. Am. Med. Assoc.*, **285**, 1992–1995.
- Moher, D., Schulz, K. F., and Altman, D. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *J. Am. Med. Assoc.*, **285**, 1987–1991.
- Moore, T. J. (1989). The cholesterol myth. *Atlantic Monthly*, Sept. 37–70.
- Moore, T. J. (1995). *Deadly Medicine*. Simon & Schuster, New York.
- Morgan, P. P. (1985). Randomized clinical trials need to be more clinical. *J. Am. Med. Assoc.*, **253**, 1782–1783.
- Morikawa, T., and Yoshida, M. (1995). A useful testing strategy in phase III trials: Combined test of superiority and test of equivalence. *J. Biopharm. Stat.*, **5**, 297–306.
- Morrison, D. F. (1976). *Multivariate Statistical Methods*, 2nd Ed. McGraw-Hill, New York.
- Moses, L. E. (1992). Statistical concepts fundamental to investigations. In *Medical Uses of Statistics*, Ed. by Bailar, J. C., and Mosteller, F. New England Journal of Medicine Books, Boston, 5–26.
- Moss, A. J., Zareba, W., Hall, J., Klein, H., Wilber, D. J., Cannom, D. S., Daubert, J. P., Higgins, S. L., Brown, M. W., and Andrews, M. L. (2002). Prophylactic implantation of a defibrillator in patients with myocardial infarction and reduced ejection fraction. *New Engl. J. Med.*, **346**, 877–883.
- Myers, R. H., and Montgomery, D. C. (1995). *Response Surface Methodology: Product and Process and Optimization with Designed Experiments*. Wiley, New York.
- Nagi, D. K., and Yudkin, J. S. (1993). Effects of Metformin on insulin resistance, risk factors for cardiovascular disease and plasminogen activator inhibitor in NIDDM subjects: A study of 2 ethnic groups. *Diabet. Care*, **16**, 621–629.

- Nam, J.-m. (1995). Sample size determination in stratified trials to establish the equivalence of two treatments. *Stat. Med.*, **14**, 2037–2049.
- Nam, J.-m. (1997). Establishing equivalence of two treatments and sample size requirements in matched-paired design. *Biometrics*, **53**, 1422–1430.
- National Institute of Neurological Disorders and Stroke rt-PA Stroke Group (1995). Tissue plasminogen activator for acute ischemic stroke. *New Engl. J. Med.*, **333**, 1581–1587.
- National Institute of Health: NIH Almanac (1981). Publication No. 81-5, Division of Public Information, Bethesda, MD.
- Nevis, S. E. (1988). Assessment of evidence from a single multicenter trial. *Proc. of Biopharmaceutical Section of the American Statistical Association*, Alexandria, VA, 43–45.
- Newcombe, R. G. (1998). Improved confidence intervals for the difference between binomial proportions based on paired data. *Stat. Med.*, **17**, 2635–2650.
- NIH Consensus Development Panel (1994). Helicobacter pylori in peptic ulcer disease. *J. Am. Med. Assoc.*, **272**, 65–69.
- Nickas, J. (1995). Adverse event data collection and reporting: A discussion of two grey areas. *Drug Info. J.*, **29**, 1247–1251.
- Northington, B. (1996). A review of issues in the collection and reporting of adverse events. *Biopharmaceut. Rep.*, **4**, 1–5.
- O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, **40**, 1079–1087.
- O'Brien, P. C., and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, **35**, 549–556.
- O'Brien, S. J., and Dean, M. (1997). In search of AIDS-resistance genes. *Sci. Am.*, **277**, 44–51.
- O'Brien, W. M. (1968). Indomethacin: A survey of clinical trials. *Clin. Pharmacol. Ther.*, **9**, 94–107.
- O'Dell, J. R., Haire, C. E., Erikson, N., Drymalski, W., Palmer, W., Eckhoff, P. J., Garwood, V., Maloley, P., Klassen, L. W., Wees, S., Klein, H., and Moore, G. F. (1996). Treatment of rheumatoid arthritis with methotrexate alone, sulfasalazine, and hydroxychloroquine, or a combination of all three treatments. *New Engl. J. Med.*, **334**, 1287–1291.
- Olkin, I. (1995). Meta-analysis: Reconciling the results of independent studies. *Stat. Med.*, **14**, 457–472.
- O'Neill, R. T. (1988a). Assessment of safety. In *Biopharmaceutical Statistics for Drug Development*, Ed. by Peace, K. Dekker, New York.
- O'Neill, R. T. (1988b). On sample sizes to estimate the protective efficacy of a vaccine. *Stat. Med.*, **7**, 1279–1288.
- O'Neill, R. T. (1993). Some FDA perspectives on data monitoring in clinical trials in drug development. *Stat. Med.*, **12**, 601–608.
- O'Quigley, J. (1999). Another look at two phase I clinical trial designs. *Stat. Med.*, **18**, 2683–2690.
- O'Quigley, J., and Chervet, S. (1991). Methods for dosing finding studies in cancer trials: a review and results of a Monte Carlo study. *Stat. Med.*, **10**, 1647–1664.
- O'Quigley, J., and Shen, L. Z. (1996). Continual reassessment method: a likelihood approach. *Biometrics*, **52**, 673–684.
- O'Quigley, J., Pepe, M., and Fisher, L. (1990). Continual reassessment method: A practical design for phase I clinical trials in cancer. *Biometrics*, **46**, 33–48.
- Pagano, M., and Halvorsen, K. T. (1981). An algorithm for finding the exact significance levels of $r \times c$ contingency tables. *J. Am. Stat. Assoc.*, **76**, 931–934.
- Park, T., and Davis, C. S. (1993). A test of the missing mechanism for repeated categorical data. *Biometrics*, **49**, 631–638.

- Patel, J. A., Reisner, B., Vizirnia, N., Owen, M., Chonmaitree, T., and Howie, V. (1995). Bacteriologic failure of amoxicillin-clavulanate in treatment of acute otitis media caused by nontypeable *haemophilus influenzae*. *J. Pediat.*, **126**, 799–806.
- Patulin Clinical Trials Committee (of the Medical Research Council) (1944). Clinical trial of Patulin in the common cold. *Lancet*, **2**, 373–375.
- Pawitan, Y., and Hallstrom, A. (1990). Statistical interim monitoring of the cardiac arrhythmia suppression trial. *Stat. Med.*, **9**, 1081–1090.
- PDR (1992). *Physicians Desk Reference*. p. 1089.
- Peace, K. E., Ed. (1987). *Biopharmaceutical Statistics for Drug Development*. Dekker, New York.
- Peace, K. E., Ed. (1990). *Statistical Issues in Drug Research and Development*. Dekker, New York.
- Peace, K. E. (1990). Response surface methodology in the development of antianginal drugs. In *Statistical Issues in Drug Research and Development*, Ed. by Peace, K. Dekker, New York, 285–301.
- Peace, K. E. (1991). One-sided or two-sided p values: Which most appropriately address the question of drug efficacy. *J. Biopharm. Stat.*, **1**, 133–138.
- Peace, K. E., Ed. (1992). *Biopharmaceutical Sequential Statistical Applications*. Dekker, New York.
- Pearson, E. S., and Hartley, H. O. (1951). Charts of the power function of the analysis of variance tests, derived from the non-central F distribution. *Biometrika*, **38**, 112–130.
- Pepe, M. S., and Cai, J. (1993). Some graphical displays and marginal regression analysis for recurrent failure times and time dependent covariates. *J. Am. Stat. Assoc.*, **88**, 811–820.
- Peterson, A. V., Mann, S. L., Kealey, K. A., and Marek, P. M. (2000). Experimental design and methods for school-based randomized trials: experience from the Hutchinson Smoking Prevention Project (HSPP). *Controlled Clin. Trials*, **21**, 144–165.
- Peterson, W. L. (1991). Drug therapy: *Helicobacter pylori* and peptic ulcer disease. *New Engl. J. Med.*, **324**, 1043–1048.
- Peto, R., and Baigent, C. (1998). Trial: The next 50 years. *Br. Med. J.*, **317**, 1170–1171.
- Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., and Smith, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. *Int. Br. J. Cancer*, **34**, 585–612.
- Petitti, D. B. (1994). *Meta-Analysis, Decision Analysis and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine*. Oxford University Press, New York.
- Petricciani, J. C. (1981). An overview of FDA, IRBs and regulations. *IRB*, **3**, 1.
- PHSRG (1989). Steering Committee of the Physician's Health Study Research Group. Final report of the aspirin component of the ongoing physicians' health study. *New Engl. J. Med.*, **321**, 129–135.
- Piaggio, G., Carroli, G., Villar J., Pinol, A., Bakketeg, L., Lumbiganon, P., Bergsjo, P., Al-Mazrou, Y., Ba'aqeel, H., Belizan, J. M., Farnot, U., and Berendes, H. (2001). Methodological considerations on the design and analysis of an equivalence stratified cluster randomization trial. *Stat. Med.*, **20**, 401–416.
- Piantadosi, S., and Liu, G. H. (1996). Improved designs for dose escalation studies using pharmacokinetic measurements. *Stat. Med.*, **15**, 1605–1618.
- Piantadosi, S., Fisher, J. D., and Grossman, S. (1998). Practical implementation of a modified continual reassessment method for dose-finding trials. *Cancer Chemother. Pharmacol.*, **41**, 429–436.
- Piantadosi, S. (1997). *Clinical Trials: A Methodologic Perspective*. Wiley, New York.
- Pierce, M., Crampton, S., Henry, D., Heifets, L., LaMarca, A., Montecalvo, M., Wormser, G. P., Jablonowski, H., Jemsek, J., Cynamon, M., Yangco, B. G., Notario, G., Craft, J. C. (1996). A randomized trial of clarithromycin as prophylaxis against disseminated *Mycobacterium avium* complex infection in patients with advanced acquired immunodeficiency syndrome. *New Engl. J. Med.*, **335**, 384–391.

- Phillips, K. F. (1990). Power of the two one-sided tests procedure in bioequivalence. *J. Pharmacokinetics & Biopharmaceutics*, **18**, 137–144.
- Pledger, G. W., and Hall, D. (1986). Active control trials: Do they address the efficacy issue? *Proc. of Biopharmaceutical Section of the American Statistical Association*, Alexandria, VA, 1–7.
- PLCO Project Team (2000). Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial. *Controlled Clin. Trials*, **21**, 273S–309S.
- Pledger, G., and Hall, D. (1990). Active control equivalence studies: Do they address the efficacy issue. In *Statistical Issues in Drug Research and Development*, Ed. by Peace, K. E. Dekker, New York, 226–238.
- PMA (1989). Issues in data monitoring and interim analysis in the pharmaceutical industry. The PMA Biostatistics and Medical Ad hoc Committee on Interim Analysis. Pharmaceutical Manufacturing Association.
- PMA Biosatistics and Medical Ad hoc Committee on Interim Analysis (1993). Interim analysis in the pharmaceutical industry. *Controlled Clin. Trials*, **14**, 160–173.
- Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, **64**, 191–199.
- Pocock, S. J. (1984). *Clinical Trials: A Practical Approach*. Wiley, New York.
- Pocock, S. J. (1990). Discussion of paper by C. B. Begg. *Biometrika*, **77**, 480–481.
- Pocock, S. J. (1997). Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. *Controlled Clin. Trials*, **18**, 530–545.
- Pocock, S. J., and Hughes, M. D. (1989). Practical problems in interim analyses with particular regard to estimation. *Controlled Clin. Trials*, **10**, 209S–221S.
- Pocock, S. J., Geller, N. L., and Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics*, **43**, 487–498.
- Pocock, S. J., O'Brien, P. C., and Fleming, T. R. (1987). A paired Prentice-Wilcoxon test for censored paired data. *Biometrics*, **43**, 169–180.
- Pocock, S. J., and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trials. *Biometrics*, **31**, 103–115.
- Powderly, W. G., Finkelstein, D. M., Feinberg, J., Frame, P., He, W., Van Der Horst, C., Koletar, S. L., Eyster, M. E., Carey, J., Waskin, H., Hooton, T. M., Hyslop, N., Spector, S., and Bozzette, S. A. (1995). A randomized trial comparing fluconazole with clotrimazole troches for the prevention of fungal infections in patients with advanced human immunodeficiency virus infection. *New Engl. J. Med.*, **332**, 700–705.
- Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika*, **65**, 167–179.
- Prentice, R. L., Williams, B. J., and Peterson, A. V. (1981). On regression analysis of multivariate failure time data. *Biometrika*, **68**, 373–379.
- Proschan, M. A., Follmann, D. A., and Geller, N. L. (1994). Monitoring multiarmed trials. *Stat. Med.*, **13**, 1441–1452.
- Randles, R. H., and Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics*. Wiley, New York.
- Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. Wiley, New York.
- Rapaport, E. (1993). GUSTO: Assessment of the preliminary results. *J. Myocardial Ischemia*, **5**, 15–24.
- Rapp, E., Pater, J. L., Willan, A., et al. (1988). Chemotherapy can prolong survival in patients with advanced non-small-lung cancer—report of a Canadian multicenter trial. *J. Clin. Oncol.*, **6**, 633–641.
- Rasmussen, D. (2000). Balancing quality versus time. *Data Basics*, **6**, 2–4.
- Ravdin, P. M., and Chamness, G. C. (1995). The c-erbB-2 proto-oncogene as a prognostic and predictive markers in breast cancer: A paradigm for the development of other macromolecular markers—a review. *Gene*, **159**, 19–27.

- Reboussin, D. M., DeMets, D. L., Kim, K. M., and Lan, K. K. G. (2000). Computation for group sequential boundaries using Lan-DeMets spending function method. *Controlled Clin. Trials*, **21**, 190–207.
- Recommendations for the treatment of hyperlipidemia in adults (1984). A joint statement of the Nutrition Committee and the Council on arteriosclerosis of the American Heart Association. *Arteriosclerosis*, **4**, 445A–468A.
- Relman, A. S. (1983). Lessons from the Darsee affair. *New Engl. J. Med.*, **308**, 1415–1417.
- Rennie, D. (1996). How to report randomized controlled trials: The CONSORT statement. *J. Am. Med. Assoc.*, **276**, 649.
- Reynolds, T. (2000). The ethics of placebo-controlled trials. *Ann. Int. Med.*, **144**, 491–492.
- Rider, P. M., O'Donnell, C., Marder, V. J., and Hennekens, C. H. (1993). Large-scale trials of thrombolytic therapy for acute myocardial infarction: GISSI-2, ISIS-3, and GUSTO-1 (Editorial). *Ann. Int. Med.*, **119**, 530–532.
- Rider, P. M., O'Donnell, C., Marder, V. J., and Hennekens, C. H. (1994). A response to “Holding GUSTO up to the light.” *Ann. Int. Med.*, **120**, 882–885.
- Ridout, M. (1991). Testing for random dropouts in repeated measurement data. *Biometrics*, **47**, 1617–1721.
- Rodda, B. E., Tsianco, M. C., Bolognese, J. A., and Kersten, M. K. (1988). Clinical development. In *Biopharmaceutical Statistics for Drug Development*, Ed. by Peace, K. E. Dekker, New York.
- Rohmel, J. (1998). Therapeutic equivalence investigations: Statistical considerations. *Stat. Med.*, **17**, 1703–1714.
- Rohmel, J., and Mansmann, U. (1999). Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority, and/or superiority. *Biomet. J.*, **41**, 149–170.
- Rosenberger, W. F. (1993). Asymptotic inference with response-adaptive treatment allocation designs. *Ann. Stat.*, **210**, 2098–2107.
- Rosenberger, W. F. (1999). Randomized play-the-winner clinical trials: Review and recommendations. *Controlled Clinical Trials*, **20**, 328–342.
- Rosen, R. C., Riley, A., Wagner, C., Osterloh, I. H., and Kirkpatrick, M. A. (1997). The International Index of Erectile Function (IIEF): A multidimensional scale for assessment of erectile dysfunction. *Urology*, **49**, 822–830.
- Rosenberger, W. F., and Lachin, J. M. (1993). The use of responsive-adaptive designs in clinical trials. *Controlled Clin. Trials*, **14**, 471–484.
- Rosner, B., and Muñoz, A. (1988). Autoregressive modelling for analysis of longitudinal data with unequally space examinations. *Stat. Med.*, **7**, 59–71.
- Rothman, K. J., and Michels, K. B. (1994). The continued unethical use of placebo controls. *New Engl. J. Med.*, **331**, 394–398.
- Rothman, K. S. (1986). *Modern Epidemiology*. Little Brown, Boston.
- Rowinsky, E. K., and Donehower, R. C. (1995). Drug therapy: Paclitaxal (taxol). *New Engl. J. Med.*, **332**, 1004–1014.
- Ruberg, S. J. (1995a). Dose response studies: I. Some design considerations. *J. Biopharm. Stat.*, **5**, 1–14.
- Ruberg, S. J. (1995b). Dose response studies: II. Analysis and interpretation. *J. Biopharm. Stat.*, **5**, 15–42.
- Rubin, B. P., Singer, S., Tsao, C., Duensing, A., Lux, M. L., Ruiz, R., Hibbard, M. K., Chen, C. J., Xiao, S., Tuveson, D. A., Demetri, G. D., Fletcher, C. D., and Fletcher, J. A. (2001). KIT activation is a ubiquitous feature of gastrointestinal stromal tumors. *Cancer Research*, **61**, 8118–8121.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

- Rubins, H. R. (1994). From clinical trials to clinical practice: Generation from participant to patient. *Controlled Clin. Trials*, **17**, 7–10.
- Rubinstein, L. V., Gail, M. H., and Santner, T. J. (1981). Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *J. Chronic Dis.*, **34**, 469–479.
- Ruskin, J. N. (1989). The Cardiac Arrhythmia Suppression Trial (CAST). *New Engl. J. Med.*, **321**, 386–388.
- Sackett, D. L. (1979). Bias in analytical research. *J. Chronic Disease*, **32**, 51–63.
- Sackett, D. L. (1989). Inference and decision at the bedside. *J. Clin. Epidemiol.*, **42**, 309–316.
- Sackett, D. L., Haynes, R. B., and Tugnell, P. (1991). *Clinical Epidemiology, A Basic Science for Clinical Medicine*. 2nd Ed., Little, Brown and Company, Boston, MA.
- Sahai, H., and Khurshid, A. (1996). Formulae and tables for the determination of sample sizes and power in clinical trials for testing difference in proportions for the two-sample design: a review. *Stat. Med.*, **15**, 1–21.
- Salsburg, D. S. (1993). The use of hazard functions in safety analysis. In *Drug Safety Assessment in Clinical Trials*. Ed. by Sogliero-Gibert, G. Dekker, New York.
- Sanford, R. L. (1994). The wonders of placebo. In *Statistics in the Pharmaceutical Industry*, Ed. by Buncher, C. R., and Tsay, J. Y. Dekker, New York.
- Sargent, D. J., and Goldberg, R. M. (2001). A flexible design for multiple armed screening trials. *Stat. Med.*, **20**, 1051–1060.
- SCDM (2000). Society for Clinical Data Management. Good Clinical Data Management Practices Committee. Good Clinical Data Management Practices, Version 1. Society for Clinical Data Management, Hillsborough, NJ.
- Scherer, J. C., and Wiltse, C. G. (1996). Adverse events: After 58 years, do we have it right yet? *Bio-pharmaceut. Rep.*, **4**, 1–5.
- Schiller, J. H., Harrington, D., Belani, C., Langer, C., Sandler, A., Krock, J., Zhu, J., and Johnson, D. H. (2002). Comparison of four Chemotherapy Regimens for Advanced Non-small-cell Lung Cancer. *New Engl. J. Med.*, **346**, 92–98.
- Schlesselman, J. J. (1982). *Case-Control Studies: Design, Conduct, Analysis*. Oxford University Press, New York.
- Schneiderman, M. A. (1967). Mouse to man: Statistical problems in bringing a drug to clinical trial. *Proc. of Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 4. University of California Press. Berkeley, CA, 855–866.
- Schoenfeld, D. (1981). Table, life; test, logmark; test, Wilcoxon: The asymptotic properties of non-parametric tests for comparing survival distributions. *Biometrika*, **68**, 316–319.
- Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioequivalence. *J. Pharmacokinetics and Biopharmaceutics*, **15**, 657–680.
- Searle, S. R. (1971). *Linear Models*. Wiley, New York.
- Seidl, L. G., Thornton, G. F., Smith, J. W., and Gluff, L. E. (1966). Studies on epidemiology of adverse drug reactions. *Bull. Hopkins Hosp.*, **119**, 299–315.
- Self, S. and Mauritsen, R. (1988). Power/sample size calculations for generalized linear models. *Biometrics*, **44**, 79–86.
- Self, S., Mauritsen, R., and Ohara, J. (1992). Power calculations for likelihood ratio tests in generalized linear models. *Biometrics*, **48**, 31–39.
- Self, S., Prentice, R., Iverson, D., Henderson, M., Thompson, D., Byar, D., Insull, W., Gorbach, S. L., Clifford, C., Goldman, S., Urban, N., Sheppard, L., and Greenwald, P. (1988). Statistical design of the women's health trial. *Controlled Clin. Trials*, **9**, 119–136.
- Segraves, R. T. (1988). Sexual side-effects of psychiatric drugs. *Int. J. Psychiat. Med.*, **18**, 243–251.

- Segraves, R. T. (1992). Sexual dysfunction complicating the treatment of depression. *J. Clin. Psychiat. Monograph*, **10**, 75–79.
- Senn, S. (1993). Inherent difficulties which active control equivalence studies. *Stat. Med.*, **12**, 2367–2375.
- SERC (1993). EGRET SIZ: sample size and power for nonlinear regression models. Reference Manual, Version 1. Statistics and Epidemiology Research Corporation.
- Seshadri, R., Figaira, F. A., Horsfall, D. J., McCaul, K., Setlur, V., and Kitchen, P. (1993). Clinical significance of HER-2/neu oncogene amplification in primary cancer. *J. Clin. Oncol.*, **11**, 1936–1942.
- Shalala, D. (2000). Protecting research subjects: What must be done? *New Engl. J. Med.*, **343**, 808–810.
- Shao, J., and Chow, S. C. (1993). Two-stage sampling with pharmaceutical applications. *Stat. Med.*, **12**, 1999–2008.
- Shao, J., and Chow, S. C. (2002). Reproducibility probability in clinical trials. *Stat. Med.*, **21**, 1727–1742.
- Shapiro, S. H., and Louis, T. A. (1983). *Clinical Trials, Issues and Approaches*. Dekker, New York.
- Shea, H. (2000). Enabling and harmonizing quality when working across multiple sites. *Data Basics*, **6**, 10–12.
- Shelton, R. C., Keller, M. B., Gelenberg, A., Dunner, D. L., Hirschfeld, R., Thase, M. E., Russell J., Lydiard, R. B., Critis-Christoph, P., Gallop, R., Todd, L., Hellerstein, D., Goodnick, P., Keitner, G., Stahl, S. M., and Halbreich, U. (2001). Effectiveness of St John's wort in major depression: A randomized controlled trial. *J. Am. Med. Assoc.*, **285**, 1978–1986.
- Shen, L. Z., and O'Quigley, J. (1996). Consistency of continual reassessment method under model misspecification. *Biometrika*, **83**, 395–405.
- Shih, J. H. (1995). Sample size calculation for complex clinical trials with survival endpoints. *Controlled Clin. Trials*, **16**, 395–407.
- Shih, W. J. (2001a). Clinical trials for drug registrations in Asian-pacific countries: proposal for a new paradigm from a statistical perspective. *Controlled Clin. Trials*, **22**, 357–366.
- Shih, W. J. (2001b). Sample size re-estimation—a journey for a decade, *Stat. Med.*, **20**, 515–518.
- Shih, W. J., Gould, A. L., and Hwang, I. K. (1989). The analysis of titration studies in phase III clinical trials. *Stat. Med.*, **8**, 583–591.
- Siegel, J. P. (2000). Equivalence and noninferiority. *Am. Heart J.*, **139**, S166–S170.
- Silliman, N. P. (1996). Analysis of subgroups in safety data. Presented at the Biopharmaceutical Section Workshop on Adverse Events, October 28–29, 1996, Bethesda, MD.
- Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clin. Trials*, **10**, 1–10.
- Simon, R. (1991). A decade of progress in statistical methodology for clinical trials. *Stat. Med.*, **10**, 1789–1817.
- Simon, R. (2000). Are placebo-controlled clinical trials ethical or needed when alternative treatment exists? *Ann. In. Med.*, **144**, 474–475.
- Simon, R., and Hall, P. (1997). Phase II trials. In *Encyclopedia of Biostatistics*, Ed. by Armitage, P. and Colton, T., Wiley, New York, 3370–3376.
- Simon, R., and Korn, E. L. (1991). Selection combinations of chemotherapeutic drugs to maximize dose intensity. *J. Biopharmaceut. Stat.*, **1**, 247–259.
- Simon, R., Freidlin B., Rubinstein, L., Arbuck, S. G., Collins, J., and Christian, M. C. (1997). Accelerated titration designs for phase I trials in Oncology. *J. Nat. Cancer Ins.*, **89**, 1138–1147.
- Simon, R., Wittes, R. E., and Ellenburg, S. S. (1985). Randomized phase II clinical trials, *Cancer Treat. Rep.*, **69**, 1375–1381.
- Sinclair, J. C. (1966). Prevention and treatment of the respiratory distress syndromes. *Pediatric. Clin. North Am.*, **13**, 711–730.

- Slamon, D. J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M., Baselga, J., and Norton, L. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpress HER2. *New Engl. J. Med.*, **344**, 792–883.
- Sleight, P. (1993). Thrombolysis after GUSTO: A European perspective. *J. Myocardial Ischemia*, **5**, 25–30.
- Slud, E. V., and Wei, L. J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *J. Am. Stat. Assoc.*, **77**, 862–868.
- Smith, A., Traganza, E., and Harrison, G. (1969). Studies on the effectiveness of antidepressant drugs. *Psychopharmacol. Bull.*, **5**(suppl. 1), 1–53.
- Smoking and Health (1964). Report of Advisory Committee to the Surgeon General of the Public Health Services. Government Printing Office, Washington, DC 235–257 (Public Health Service Publication no. 1103).
- Snapinn, S. M. (2001). Alternative for discounting historical data in the analysis of noninferiority trial. *ICSA Bulletin*, January, 29–33.
- Snedecor, G. W., and Cochran, W. G. (1980). *Statistical Methods*, 7th Ed. Iowa State University, Ames.
- Spilker, B. (1991). *Guide to Clinical Trials*. Raven Press, New York.
- Spilker, B., and Schoenfelder, J. (1984). *Data Collection Forms for Clinical Trials*. Raven Press, New York.
- Spriet, A. and Dupin-Spriet, T. (1996). *Good Practice of Clinical Drug Trials*. Karger, S. Karger AG, Medical & Scientific Publication, Basel.
- Stampfer, M. J., Willett, W. C., and Colditz, G. A. (1985). A prospective study of postmenopause estrogen therapy and coronary heart disease. *New Engl. J. Med.*, **313**, 1044–1049.
- Stanley, B. (1988). An integration of ethical and clinical considerations in the use of placebos. *Psychopharmacol. Bull.*, **24**, 18–20.
- Staszewski, S., Keiser, P., Montaner, J., Raffi, F., Gathe, J., Brotas, V., Hicks, C., Hammer, S. M., Cooper, D., Johnson, M., Tortell, S., Cutrell, A., Thorborn, D., Isaacs, R., Hetherington, S., Steel, H., and Spreen, W. CNAAB3005 International Study Team. (2001) Abacavir-lamuvidine-zidovudine vs. indinavir-lamuvidine-zidovudine in antiretroviral-naïve-HIV-infected adults: a randomized equivalence trials. *J. Am. Med. Assoc.*, **285**, 1155–1163.
- Steinbrook, R. (2002a). Improving protection for research subjects. *New Engl. J. Med.*, **346**, 716–720.
- Steinbrook, R. (2002b). Improving protection for research subjects. *New Engl. J. Med.*, **346**, 1425–1430.
- Steward, R. B., and Cluff, L. E. (1972). A review of medication errors and compliance in ambulant patients. *Clin. Pharmacol. Ther.*, **13**, 463–468.
- Stolley, P. D. and Storm, B. L. (1986). Sample size calculations for clinical pharmacology studies. *Clin. Pharmacol. Ther.*, **27**, 489–490.
- Storer, B. E. (1989). Design and analysis of phase I trials. *Biometrics*, **45**, 925–937.
- Storer, B. E. (1993). Small-sample confidence sets for the MTD in a phase I clinical trial. *Biometrics*, **49**, 1117–1125.
- Storer, B. E. (1997). Phase I trials. In *Encyclopedia of Biostatistics*, Ed. by Armitage, P., and Colton, T., Wiley, New York, 3365–3370.
- Storer, B. E. (2001). An evaluation of phase I clinical trial designs in the continuous dose-response setting. *Stat. Med.*, **20**, 2399–2408.
- Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, **42**, 412–416.
- Suissa, S., and Shuster, J. J. (1991). The 2×2 matched-pairs trial: Exact unconditional design and analysis. *Biometrics*, **47**, 361–372.

- Tamura, R. N., Faries, D. E., Andersen, J. S., and Heiligenstein (1994). A case study of an adaptive clinical trial in the treatment of out-patients with depression disorder. *J. Am. Stat. Assoc.*, **89**, 768–776.
- Tang, D.-I., Geller, N. L., and Pocock, S. J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics*, **49**, 23–30.
- Tango, Y. (1998). Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Stat. Med.*, **17**, 891–908.
- Tarone, R. E., and Ware, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika*, **64**, 156–160.
- Taves, D. R. (1974). Minimization: A new method of assigning patients to treatment and control groups. *Clin. Pharmacol. Ther.*, **15**, 443–453.
- Temple, R. (1982). Government viewpoint of clinical trials. *Drug Info. J.*, **16**, 10–17.
- Temple, R. (1983). Difficulties in evaluating positive control trials. *Proc. of Biopharmaceutical Section of the American Statistical Association*, Alexandria, VA, 1–7.
- Temple, R. (1993). Trends in pharmaceutical development. *Drug Info. J.*, **27**, 355–366.
- Temple, R. (1996). Problems in interpreting active control equivalence trials. *Accountability in Research*, **4**, 267–275.
- Temple, R. (1997). When are clinical trials of a given agent vs. placebo no longer appropriate or feasible? *Controlled Clin. Trials*, **18**, 613–620.
- Temple, R., and Ellenburg, S. S. (2000). Placebo—Controlled Trials and Active-Controlled Trials in the Evaluation of New Treatments, Part I: Ethical and Scientific Issues. *Ann. Int. Med.*, **133**, 455–463.
- Tessman, D. K., Gipson, B., and Levins, M. (1994). Cooperative fast-track development: The fludara story. *Appl. Clin. Trials*, **3**, 55–62.
- Testa, M. A., Anderson, R. B., Nackley, J. F., Hollenberg, N. K., and the Quality-of-Life Hypertension Study Group (1993). Quality of life and antihypertensive therapy in men—A comparison of Captopril with Enalapril. *New Engl. J. Med.*, **328**, 907–913.
- Thall, P. F., and Simon, R. (1990). Incorporating historical control data in planning phase II clinical trials. *Stat. Med.*, **9**, 215–228.
- Thall, P. F., and Simon, P. (1994a). Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics*, **50**, 337–349.
- Thall, P. F., and Simon, P. (1994b). Bayesian approach to establishing sample size and monitoring criteria for phase II clinical trials. *Controlled Clin. Trials*, **15**, 463–481.
- Thall, P. E., Lee, J. J., Tseng, C. H., and Estey, E. H. (1999). Accrual strategies for phase I trials with delayed patient outcome. *Stat. Med.*, **18**, 1155–1169.
- Thall, P. F., Simon, R. M., and Estey, E. H. (1995). Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes, *Stat. Med.*, **14**, 357–379.
- Thall, P. F., Simon, R. M., and Estey, E. H. (1996). New statistical strategy for monitoring safety and efficacy in single-arm clinical trials, *J. Clin. Oncol.*, **14**, 296–303.
- Thall, P. E., Sung, H. G., and Choudhury, A. (2001). Dose-finding based on feasibility and toxicity in T-cell infusion trials. *Biometrics*, **57**, 914–921.
- The Economist*. (2002). Pharmaceutical mercky prospects. 364, July 13–19, London, U.K.
- The Expert Panel (1988). Report of the National Cholesterol Education Program Expert Panel on Detection, Evaluation and Treatment of High Blood Cholesterol in Adults. *Archiv. Int. Med.*, **148**, 36–69.
- The Global Use of Strategies to Open Occluded Coronary Arteries (GUSTO) IIb Investigators (1996). A comparison of recombinant hirubin with heparin for the treatment of acute coronary syndromes. *New Engl. J. Med.*, **335**, 775–782.

- The West of Scotland Coronary Prevention Study Group (1995). Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. *New Engl. J. Med.*, **333**, 1301–1307.
- Therasse, P., Arbuck, S. G., Eisenhauer, E. A., Wanders, J., Kaplan, R. S., Rubinstein, L., Verweij, J., Van Glabbeke, M., van Oosterom, A. T., Christian, M. C., and Gwyther, S. G. (2000). New guidelines to evaluate the response to treatment in solid tumors. *J. Nat. Cancer Inst.*, **92**, 205–216.
- Therneau, T. M., Wieand, H. S., and Chang, M. N. (1990). Optimal designs for a grouped sequential binomial trial. *Biometrics*, **46**, 771–781.
- Ting, N., Burdick, R. K., Graybill, F. A., Jeyaratnam, S., and Lu, T. F. C. (1990). Confidence intervals on linear combinations of variance components that are unrestricted in sign. *Journal of Statistical Computation and Simulation*, **35**, 135–143.
- Tremmel, L. (1996). Describing risk in long-term clinical trials. *Biopharmaceut. Rep.*, **4**, 5–8.
- Tsai, K. T., and Patel, H. I. (1992). Exploratory data analysis of a multicenter trial. *Proc. of Biopharmaceutical Section of the American Statistical Association*, Alexandria, VA, 56–61.
- Tsiatis, A. A., Rosner, G. L., and Mehta, C. R. (1984). Exact confidence interval following a group sequential test. *Biometrics*, **40**, 797–803.
- Tsong, Y., Wang, S. J., Cui, L., and Hung, J. H. M. (2001). Placebo control, historical control, and active control trials, *ICSA Bulletin*, January, 36–39.
- Tugwell, P., Pincus, T., Yocum, D., Stein, M., Gluck, O., Kraag, G., McKendry, R., Tesser, J., Baker, P., and Wells, G., for the Methotrexate-Cyclosporine Combination Study Group (1995). *New Engl. J. Med.*, **333**, 137–141.
- Tystrup, N., Lachin, J. M., and Juhl, E. (1982). *The Randomized Clinical Trials and Therapeutic Decisions*. Dekker, New York.
- USP/NF (2002). The United States Pharmacopeia 26 and the National Formulary XVIII. The United States Pharmacopeial Convention, Rockville, MD.
- van Elteren, P. H. (1960). On the combinations of independent two-sample tests of Wilcoxon. *Bull. Int. Stat. Inst.*, **37**, 351–361.
- Veteran's Administration Cooperative Urological Research Group (1967). Treatment and survival of patients with cancer of the prostate. *Surg. Gynecol. Obstet.*, **124**, 1011–1017.
- Vonesh and Chinchilli, V. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measures*. Dekker, New York.
- Walsh, J. H., and Peterson, W. L. (1995). Drug therapy: The treatment of *Helicobacter pylori* infection in the management of peptic ulcer disease. *New Engl. J. Med.*, **333**, 984–991.
- Wang, M. C., and Chang, S. H. (1999). Nonparametric estimation of a recurrent survival function. *The Journal of the American Statistical Association*, **94**, 146–153.
- Wang, M. C., and Chen, Y. Q. (2000). Nonparametric and semiparametric trend analysis for stratified recurrent times. *Biometrics*, **56**, 789–794.
- Wang, M. C., Qin, J., and Chiang, C. T. (2001). Analyzing recurrent event data with informative censoring. *The Journal of the American Statistical Association*, **96**, 1057–1065.
- Wang, S. J., and Hung, H. M. J. (2003). Assessing treatment efficacy in noninferiority trials, *Controlled Clin. Trials*, **24**, 147–155.
- Wang, S. J., and Hung, J. H. M. (2003). TACT method for noninferiority testing in active controlled trials. *Stat. Med.*, **22**, 227–238.
- Wang, S. J., Hung, J. H. M., and Tsong, Y. (2002). Utility and pitfalls of some statistical methods in active control trials. *Controlled Clin. Trials*, **23**, 15–28.
- Wang, S. G. and Chow, S. C. (1995). *Advanced Linear Models*. Dekker, New York.
- Wang, W., Hsuan, F., and Chow, S. C. (1996). Patient compliance and the fluctuation of the serum drug concentration. *Stat. Med.*, **15**, 659–669.

- Ward, D. E., and Camm, A. J. (1993). Dangerous ventricular arrhythmia: Can we predict drug efficacy. *New Engl. J. Med.*, **329**, 498–499.
- Ware, J. H. (1989). Investigating therapies of potentially great benefit: ECMO. *Stat. Sci.*, **4**, 298–340.
- Ware, J. H., and Antman, E. M. (1997). Equivalence trials. *New Engl. J. Med.*, **337**, 1159–1161.
- Ware, J. E., Jr., Kosinski, M., and Keller, S. D. (1994). *SF-36 Physical and Mental Health Summary Scales: A User's Manual*. Health Assessment Lab., New England Medical Center, Boston, MA.
- Ware, J., Lipsitz, S., and Speizer, F. E. (1988). Issue in the analysis of repeated categorical outcomes. *Stat. Med.*, **7**, 95–107.
- Ware, J. H., Mosteller, F., Delgado, F., Donnelly, C., and Ingelfinger, J. A. (1992). P Values, Chapter 10 in *Medical Uses of Statistics*, 2nd Ed. Edited by J. C. Bailar III and F. Mosteller, NEJM Books, Boston, MA.
- Warren, J. R. (1982). Unidentified curved bacilli on gastric epithelium in active chronic gastritis. *Lancet*, 1273.
- Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439–447.
- Wei, L. J. (1977). A class of designs for sequential clinical trials. *J. Am. Stat. Assoc.*, **72**, 382–386.
- Wei, L. J. (1978). The adaptive biased-coin design for sequential experiments. *Ann. Stat.*, **9**, 92–100.
- Wei, L. J. (1988). Exact two-sample permutation tests based on the randomized play-the-winner rule. *Biometrika*, **75**, 603–606.
- Wei, L. J. (1990). Discussion of paper by C. B. Begg, *Biometrika*, **77**, 476–477.
- Wei, L. J., and Durham, S. (1978). The randomized play-the-winner rule in medical trials. *J. Am. Stat. Assoc.*, **73**, 840–843.
- Wei, L. J., and Glidden, D. V. (1997). An overview of statistical methods for multiple incomplete failure time data in clinical trials. *Stat. Med.*, **16**, 833–839.
- Wei, L. J., and Lachin, J. M. (1988). Properties of the urn randomization in clinical trials. *Controlled Clin. Trials*, **9**, 345–364.
- Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distribution. *J. Am. Stat. Assoc.*, **84**, 1065–1073.
- Wei, L. J., Smythe, R. T., Lin, D. Y., and Park, T. S. (1990). Statistical inference with data-dependent treatment allocation rules. *J. Am. Stat. Assoc.*, **73**, 840–843.
- Weinshilboum, R. (2003). Inheritance and drug response. *New Engl. J. Med.*, **348**, 529–537.
- Weiss, R. B., Gill, G. G., and Hudis, C. A. (2001). An on-site audit of the South African trial of high-dose chemotherapy for metastatic breast cancer and associated publications. *J. Clin. Oncol.*, **19**, 2771–2777.
- Weintraub, M., and Calimlim, J. F. (1994). Selecting patients for a clinical trial. In *Statistics in the Pharmaceutical Industry*, Ed. by Buncher, C. R., and Tsay, J. Y. Dekker, New York.
- Wellek, S. (1993). A log-rank test for equivalence of two survivor function. *Biometrics*, **49**, 877–881.
- Weschler, H., Grosser, G. H., and Greenblatt, M. (1965). Research evaluating antidepressant medications on hospitalized mental patients: A survey of published reports during a five-year period. *J. Nervous Mental Dis.*, **141**, 21–239.
- Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *J. Pharmaceut. Sci.*, **61**, 1340–1341.
- Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics*, **32**, 741–744.
- White, S. J., and Freedman, L. S. (1978). Allocation of patients to treatment groups in a controlled clinical study. *Br. J. of Cancer*, **37**, 849–857.

- WHO (1979). WHO Handbook for reporting results of cancer treatment, World Health Organization Offset Publication No. 48, Geneva, Switzerland.
- Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*, Rev. 2nd Ed. John Wiley, Chichester, England.
- Wiens, B. L. (2002). Choosing an equivalence limit for noninferiority or equivalence studies. *Controlled Clin. Trials*, **23**, 2–14.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, **1**, 80–83.
- Williams, D. A. (1971). A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*, **27**, 103–118.
- Williams, D. A. (1972). The comparison of several dose levels with a zero dose control. *Biometrics*, **28**, 519–531.
- Williams, E. J. (1949). Experimental designs balanced for the residual effects of treatment. *Austral. J. Sci. Res.*, **2**, 149–168.
- Williams, G. W., Davis, R. L., Geston, A. J., Gould, L., Hwang, I. K., Mathews, Shih, W. J., Snapinn, S. M., and Walton-Bowen, K. L. (1993). Monitoring of clinical trials and interim analyses for a drug sponsor's point of view. *Stat. Med.*, **12**, 481–492.
- Wilson, P. W. F., Garrison, R. J., and Castelli, W. P. (1985). Post-menopause estrogen use, cigarette smoking and cardiovascular morbidity in women over 50: The Framingham study. *New Engl. J. Med.*, **313**, 1038–1043.
- Winkler, C., Modi, W., Smith, M. W., Nelson, G. W., Wu, X., Carrington, M., Dean, M., Honjo, T., Tashiro, K., Yabe, D., Buchbinder, S., Vittinghoff, E., Goedert, J. J., O'Brien, T. R., Jacobson, L. P., Detels, R., Donfield, S., Willoughby, A., Gomperts, E., Vlahov, D., Phair, J., and O'Brien, S. J. (1998). Genetic restriction of AIDS pathogenesis by an SDF-1 chemokine gene variant, *Science*, **279**, 389–393.
- Wittes, J. (1994). Introduction: From clinical trials to clinical practice—four papers from a plenary session. *Controlled Clin. Trials*, **15**, 5–6.
- Wittes, J. (1996). A statistical perspective on adverse event reporting in clinical trials. *Biopharmaceut. Rep.*, **4**, 5–10.
- Wittes, J. (2001). Active-control trials: a linguistic problem. *ICSA Bulletin*, January, 9–40.
- Wolf, S. (1950). Affects of suggestion and conditioning on the action of chemical agents in human subjects—The pharmacology of placebos. *J. Clin. Invest.*, **29**, 100–109.
- Wooding, W. M. (1994). *Planning Pharmaceutical Clinical Trials: Basic Statistical Principles*. Wiley, New York.
- Woollen, S. W. (2000a). Patients missue and investigator fraud in clinical trials: what can be done? Part I, an invited presentation at the 2000 Annual Meeting of Drug Information Association, June 14, San Diego, CA.
- Woollen, S. W. (2000b). Detecting and handling scientific misconduct and persistent noncompliance—the FDA's perspective. An invited presentation at the 2000 Annual Meeting of Drug Information Association, June 14, San Diego, CA.
- Women's Health Initiative Study Group (1998). Design of the Women's Health Initiative Clinical Trial and Observational Study. *Controlled Clin. Trials*, **19**, 61–109.
- Writing Group for the Women's Health Initiative (2002). Risks and Benefits of Estrogen plus progestin in healthy postmenopausal women: Principal results from the Women's Health Initiative Randomized Controlled Trial. *J. Am. Med. Assoc.*, **288**, 321–333.
- Wu, M., Fisher, M., and DeMets, D. (1980). Sample sizes of long-term medical trials with time-dependent noncompliance and event rates. *Controlled Clin. Trials*, **1**, 109–121.
- Yam, A. L. (2003). Good programming practice. In *Encyclopedia of Biopharmaceutical Statistics*, 2nd Ed. by Chow, S.C. Dekker, New York.

- Yamuah, L. K. (2001). The role of data management in health: A case study in developing countries. *Drug Info. J.*, **35**, 707–712.
- Yateman, N. A. and Skene, A. M. (1992). Sample sizes for proportional hazards survival studies with arbitrary patient entry and loss to follow-up distributions, *Stat. Med.*, **11**, 1103–1113.
- Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. *J. Am. Stat. Assoc.*, **29**, 51–66.
- Yates, F., and Cochran, W. G. (1938). The analysis of groups of experiments. *J. Agri. Sci.*, **28**, 556–580.
- Yusuf, S., Peto, R., Lewis, J., Collins, R., and Sleight, P. (1985). Beta blocker during and after myocardial infarction: an overview of the randomized trials. *Prog. Cardiovas. Dis.*, **27**, 335–371.
- Zeger, S. L., and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **44**, 1825–1829.
- Zeger, S. L., Liang, K. Y., and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.
- Zeger, S. L., and Liang, K. Y. (1992). An overview of methods for the analysis of longitudinal data. *Stat. Med.*, **11**, 1825–1839.
- Zeger, S. L., and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, **44**, 1019–1031.
- Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *J. Am. Stat. Assoc.*, **64**, 131–146.
- Zohar S., and Chevret, S. (2001). The continual reassessment method: Comparison of Bayesian stopping rules for dose-ranging studies. *Stat. Med.*, **20**, 2827–2843.
- Zoloft (1992). Zoloft (sertraline hydrochloride) tablets, package insert, Pfizer Inc. Groton, CT.

1

INTRODUCTION

1.1 WHAT ARE CLINICAL TRIALS?

Clinical trials are clinical investigations. They have evolved with different meanings by different individuals and organizations at different times. For example, Meinert (1986) indicates that a clinical trial is a research activity that involves administration of a test treatment to some *experimental unit* in order to evaluate the treatment. Meinert (1986) also defines a clinical trial as a planned experiment designed to assess the efficacy of a treatment in humans by comparing the outcomes in a group of patients treated with the test treatment with those observed in a comparable group of patients receiving a control treatment, where patients in both groups are enrolled, treated, and followed over the same time period. This definition indicates that a clinical trial is used to evaluate the effectiveness of a treatment. Piantadosi (1997) simply defined a clinical trial as an experimental testing medical treatment on human subject. On the other hand, Spilker (1991) considers clinical trials as a subset of clinical studies that evaluate investigational medicines in phases I, II, and III, the clinical studies being the class of all scientific approaches to evaluate medical disease preventions, diagnostic techniques, and treatments. This definition is somewhat narrow in the sense that it restricts to the clinical investigation conducted by pharmaceutical companies during various stages of clinical development of pharmaceutical entities which are intended for marketing approval. The Code of Federal Regulations (CFR) defines a clinical trial as the clinical investigation of a drug that is administered or dispensed to, or used involving one or more human subjects (21 CFR 312.3). Three important key words in these definitions of clinical trials are *experimental unit*, *treatment*, and *evaluation* of the treatment.

Experimental Unit

An *experimental unit* is usually referred to as a subject from a targeted population under study. Therefore the experimental unit is usually used to specify the intended study population to which the results of the study are inferred. For example, the intended population could be patients with certain diseases at certain stages or healthy human subjects. In practice, although a majority of clinical trials are usually conducted in patients to evaluate certain test treatments, it is not uncommon that some clinical trials may involve healthy human subjects. For example, at very early phase trials of clinical development, initial investigation of a new pharmaceutical entity may only involve a small number of healthy subjects, say fewer than 30. Large primary prevention trials are often conducted with healthy human subjects with size in tens of thousand subjects. See, for example, *Physician's Health Study* (PHSRG, 1988), *Helsinki Health Study* (Frick et al., 1987), and *Women Health Trial* (Self et al., 1988).

Treatment

In clinical trials a *treatment* can be a placebo or any combinations of a new pharmaceutical identity (e.g., a compound or drug), a new diet, a surgical procedure, a diagnostic test, a medical device, a health education program, or no treatment. For example, in the *Physician's Health Study*, one treatment arm is a combination of low-dose aspirin and beta carotene. Other examples include lumpectomy, radiotherapy, and chemotherapy as a combination of surgical procedure and drug therapy for breast cancer; magnetic resonance imaging (MRI) with a contrast imaging agent as a combination of diagnostic test and a drug for enhancement of diagnostic enhancement; or a class III antiarrhythmic agent and an implanted cardioverter defibrillator as a combination of a drug and a medical device for treatment of patients with ventricular arrhythmia. As a result, a *treatment* is any intervention to be evaluated in human subjects regardless that it is a new intervention to be tested or serves as a referenced control group for comparison.

Evaluation

In his definition of clinical trials, Meinert (1986) emphasizes the *evaluation* of efficacy of a test treatment. It, however, should be noted that the assessment of safety of an intervention such as adverse experiences, elevation of certain laboratory parameters, or change in findings of physical examination after administration of the treatment is at least as important as that of efficacy. Recently, in addition to the traditional evaluation of effectiveness and safety of a test treatment, clinical trials are also designed to assess quality of life, pharmacogenomics, and pharmacoeconomics such as cost-minimization, cost-effectiveness, and cost-benefit analyses to human subjects associated with the treatment under study. It is therefore recommended that clinical trials should not only evaluate the effectiveness and safety of the treatment but also assess quality of life, impact of genetic factors, pharmacoeconomics, and outcomes research associated with the treatment.

Throughout this book we will define a clinical trial as a clinical investigation in which treatments are administered, dispensed, or used involving one or more human subjects for evaluation of the treatment. By this definition, the experimental units are human subjects either with a pre-existing disease under study or healthy. Unless otherwise specified, clinical trials in this book are referred to as all clinical investigations in human subjects that may be conducted by

pharmaceutical companies, clinical research organizations such as the U.S. National Institutes of Health (NIH), university hospitals, or any other medical research centers.

1.2 HISTORY OF CLINICAL TRIALS

We humans since our early days on earth have been seeking or trying to identify some interventions, whether they be a procedure or a drug, to remedy ailments that inflict ourselves and our loved ones. In this century the explosion of modern and advanced science and technology has led to many successful discoveries of promising treatments such as new medicines. Over the years there has been a tremendous need for clinical investigations of these newly discovered and promising medicines. In parallel, different laws have been enacted and regulations imposed at different times to ensure that the discovered treatments are effective and safe. The purpose for imposing regulations on the evaluation and approval of treatments is to minimize potential risks that they may have for human subjects, especially for those treatments whose efficacy and safety are unknown or are still under investigation.

In 1906, the United States Congress passed the *Pure Food and Drug Act*. The purpose of this act is to prevent misbranding and adulteration of food and drugs. However, the scope of this act is rather limited. No preclearance of drugs is required. Moreover the act does not give the government any authority to inspect food and drugs. Since the act does not regulate the claims made for a product, the Sherley Amendment to the act was passed in 1912 to prohibit labeling medicines with false and fraudulent claims. In 1931, the U.S. Food and Drug Administration (FDA) was formed. The provisions of the FDA are intended to ensure that (1) food is safe and wholesome, (2) drugs, biological products, and medical devices are safe and effective, (3) cosmetics are unadulterated, (4) the use of radiological products does not result in unnecessary exposure to radiation, and (5) all of these products are honestly and informatively labeled (Fairweather, 1994).

The concept of testing marketed drugs in human subjects did not become a public issue until the Elixir Sulfanilamide disaster occurred in the late 1930s. The disaster was a safety concern of a liquid formulation of a sulfa drug that caused more than 100 deaths. This drug had never been tested in humans before its marketing. This safety concern led to the pass of the *Federal Food, Drug and Cosmetic Act* (FD&C Act) in 1938. The FD&C Act extended its coverage to cosmetics and therapeutic devices. More important, the FD&C Act requires the pharmaceutical companies to submit full reports of investigations regarding the safety of new drugs. In 1962, a significant Kefauver-Harris Drug Amendment to the FD&C Act was passed. The Kefauver-Harris Amendment not only strengthened the safety requirements for new drugs but also established an efficacy requirement for new drugs for the first time. In 1984, the Congress passed the *Price Competition and Patent Term Restoration Act* to provide for increased patent protection to compensate for patent life lost during the approval process. Based on this act, the FDA was also authorized to approve generic drugs only based on bioavailability and bioequivalence trials on healthy male subjects. It should be noted that the FDA also has the authority for designation of prescription drugs or over-the counter drugs. In the United States, on average, it will take a pharmaceutical company about 10 to 12 years for development of a promising pharmaceutical entity with an average cost between \$350 millions to \$450 millions US. Drug development is a lengthy and costly process. This lengthy process is necessary to ensure the safety and efficacy of the drug product under investigation. On average, it may take more than two years for regulatory authorities such as the FDA to complete the review of

the new drug applications submitted by the sponsors. This lengthy review process might be due to limited resources available at the regulatory agency. As indicated by the U.S. FDA, they will be able to improve the review process of new drug applications if additional resources are available. As a result, in 1992, the U.S. Congress passed the *Prescription Drug User Fee Act* (PDUFA), which authorizes the FDA to utilize the so-called *user fee* financed by the pharmaceutical industry to provide additional resources for the FDA's programs for development of drug and biologic products. From 1992 to 1997, this program has enabled the FDA to reduce the average time required for review of a new drug application from 30 months to 15 months. In 1997, the U.S. Congress also passed the *Food and Drug Administration Modernization Act* (FDAMA) to enhance the FDA's missions and its operations for the increasing technological, trade, and public health complexities in the 21st Century by reforming the regulation of food, drugs, devices, biologic products, and cosmetics.

The concept of randomization in clinical trials was not adapted until the early 1920s (Fisher and Mackenzie, 1923). Amerson et al. (1931) first considered randomization of patients to treatments in clinical trials to reduce potential bias and consequently to increase statistical power for detection of a clinically important difference. At the same time a Committee on Clinical Trials was formed by the Medical Research Council of the Great Britain (Medical Research Council, 1931) to promulgate good clinical practice by developing guidelines governing the conduct of clinical studies from which data will be used to support application for marketing approval. In 1937, the NIH awarded its first research grant in clinical trial. At the same time the U.S. National Cancer Institute (NCI) was also formed to enhance clinical research in the area of cancer. In 1944, the first publication of results from a multicenter trial appeared in *Lancet* (Patulin Clinical Trials Committee, 1944). Table 1.2.1 provides a chronic accounts of historical events for both clinical trials and the associated regulations for treatments intended for marketing approval. Table 1.2.1 reveals that the advance of clinical trials goes hand in hand with the development of regulations.

Oklin (1995) indicated that there are at least 8,000 randomized controlled clinical trials conducted each year whose size can include as many as 100,000 subjects. As more clinical trials are conducted worldwide each year, new service organization and/or companies have emerged to provide information and resources for the conduct of clinical trials. Table 1.2.2 provides a summary of resources available for clinical trials from a web-based clinical trial listing service called CenterWatch.[®] These trials are usually sponsored by the pharmaceutical industry, government agencies, clinical research institutions, or more recently a third party such as health maintenance organizations (HMO) or insurance companies. In recent years clinical trials conducted by the pharmaceutical industry for marketing approval have become more extensive. However, the sizes of clinical trials funded by other organizations are even larger. The trials conducted by the pharmaceutical industry are mainly for the purpose of registration for marketing approval. Therefore, they follow a rigorously clinical development plan which is usually carried out in phases (e.g., phases I, II, and III trials, which will be discussed later in this chapter) that progress from very tightly controlled dosing of a small number of normal subjects to less tightly controlled studies involving large number of patients.

According to *USA Today* (Feb. 3, 1993), the average time that a pharmaceutical company spends getting a drug to market is 12 years and 8 months. Of this figure, six years and 8 months are spent in clinical trials to obtain the required information for market registration. The FDA review takes 2 years and 6 months. As a result of PDUFA, the review time at the U.S. FDA has been reduced considerably. Table 1.2.3 provides a summary of median review time at the Center for Drug Review and Research (CDER) at the U.S. FDA in 2001.

Table 1.2.1 Significant Historical Events in Clinical Trials and Regulations

Year	Clinical Trials	Regulations
1906		Pure Food and Drug Act (Dr. Harvey Wiley)
1912		Sherley Amendment
1923	First randomization to experiments (Fisher and Mackenzie, 1923)	
1931	First randomization of patients to treatments in clinical trials <i>(Amberson, et al., 1931)</i>	Formation of U.S. Food and Drug Administration
	Committee on clinical trials by the Medical Research Council of Great Britain (Medical Research Council, 1931)	
1937	Formation of National Cancer Institute and First Research Grant by National Institutes of Health (National Institutes of Health, 1981)	
1938		U.S. Federal Food, Drug and Cosmetic Act (Dr. R. Tugwell)
1944	First publication of results from a multicenter trial (Patum Clinical Trial Committee, 1944)	
1952	Publication of <i>Elementary Medical Statistics</i> (Mainland, 1952)	FDA makes designation of Prescription Drug or OTC Amendment to the U.S. Food, Drug, and Cosmetic Act
1962	Publication of <i>Statistical Methods in Clinical and Preventive Medicine</i> (Hill, 1962)	
1966		Mandated creation of the local boards (IRB) for Funding by U.S. Public Health Service
1976		Medical Device Amendment to the U.S. Food, Drug Cosmetic Act (1976)
1977		Publications of <i>General Considerations for Clinical Evaluation of Drugs</i> (HEW (FDA), 1977)
1984		Drug Price Competition and Patent Term Restoration Act (Waxman and Hatch, 1984)
1985		NDA rewrite
1988		Publication of <i>Guidelines for the Format and Content of the Clinical and Statistical Section of an Application</i> (FDA, 1988)
1990		Publication of <i>Good Clinical Practice for Trials on Medicinal Products in the European Community</i> (EC Commission, 1990)
1987		Treatment IND (FDA, 1987)
1992		Parallel track and accelerated approval (FDA, 1992)
		Prescription Drug User Fee Act
1997		Publication of <i>Good Clinical Practice: Consolidated Guidelines</i> (ICH, 1996)
		U.S. FDA Modernization Act

Table 1.2.2 Summary of Resources for Clinical Trials

Description	Resources
Number of Clinical Trials	41,000
Clinical Investigators	25,000
Academic Clinical Research Center	600
Pharmaceutical, Biotechnology, and Medical Device Companies	275
Contract Research Organization (CRO)	250
Companies Provides Services to Clinical Trials	130
Financial and Investment Professionals for Clinical Trials	100

Source: CenterWatch® Clinical Trials Listing Service (<http://www.centerwatch.com>).

For example, for the 10 drugs receiving priority status, the median review time is only 6 months. The median overall approval time is 14 months. However, it is not surprising that new molecular entities require about more than 7 months to review. This lengthy clinical development process is necessary to assure the efficacy and safety of the drug product. As a result, this lengthy development period sometimes does not allow the access of promising drugs or therapies to subjects with serious or life-threatening illnesses. Kessler and Feiden (1995) point out that the FDA may permit promising drugs or therapies currently under investigation to be available to patients with serious or life-threatening diseases under the so-called *treatment IND* in 1987. The *Parallel Track Regulations* in 1992 allow promising therapies for serious or life-threatening diseases to become available with considerably fewer data than required for approval. In the same year, the FDA published the regulations for the *Accelerated Approval* based only on surrogate endpoints to accelerate the approval process for promising drugs or therapies indicated for life-threatening diseases.

The size of trials conducted by the pharmaceutical industry can be as small as a dozen subjects for the phase I trial in human, or it can be as large as a few thousands for support of approval of ticlopidine for stroke prevention (Temple, 1993). The design of the trial can be very simple as the single-arm trial with no control group, or it can be very complicated as a 12-group factorial design for the evaluation of the dose responses of combination drugs. Temple (1993) points out that information accumulated from previous experience in the database of preapproval New Drug Application (NDA) or Product License Application (PLA) can range from a few hundred subjects (e.g., contrast imaging agents) to four or five thousand subjects (antidepressants or antihypertensives, antibiotics, etc.).

When the safety profile and mechanism of action for the efficacy of a new drug or therapy are well established, probably after its approval, a simple but large confirmatory trial is usually conducted to validate the safety and effectiveness of the new drug or therapy. This

Table 1.2.3 Summary of Median Review Time at CDER of the U.S. FDA in 2001

Number of Approved Drugs	Median Review Time in Months
66	14
NME (24)	19
Priority status (10)	6
Standard status (56)	12

Source: FDA talk paper on January 25, 2002 at www.fda.gov.

NME = New Molecular Entities.

kind of trial is large in the sense that there are relaxed the entrance criteria to enroll a large number of subjects (e.g., tens of thousands) with various characteristics and care settings. The purpose of this kind of trial is to increase the exposure of a new drug or therapy to more subjects with the indicated diseases. For example, the first *Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries Trial* (GUSTO I, 1993) enrolled over 41,000 subjects in 1,081 hospitals from 15 countries while in the *Physician's Health Study* funded by the NIH over 22,000 physicians were randomized to one of four arms in the trial. In addition, these trials usually follow subjects for a much longer period of time than most trials for marketing approval. For example, *Helsinki Heart Study* followed a cohort over 4,000 middle-aged men with dyslipidemia for five years (Frick et al., 1987). The recent *Prostate Cancer Prevention Trial* (PCPT) plans to follow 18,000 healthy men over age 55 for 7 years (Feigl et al., 1995). Such trials are simple in the sense that only few important data are collected from each subject. Because the sizes of these trials are considerably large, they can detect a relatively small yet important and valuable treatment effects that previous smaller studies failed to detect. Sometimes, public funded clinical trials can also be used as a basis for approval of certain indications. An example is the combined therapy of leuprolide with flutamide for patients with disseminated, previously untreated D₂ stage prostate cancer. Approval of flutamide was based on a study funded by NCI.

On the other hand, health care providers such as HMO or insurance companies will be more interested in providing funding for rigorous clinical trials to evaluate not only efficacy and safety of therapies but also quality of life, pharmacoeconomics, and outcomes. The purpose of this kind of clinical trial is to study the cost associated with the health care provided. The concept is to minimize the cost with the optimal therapeutic effect under the same quality of health care. Temple (1993) points out that from the results of the study of *Systolic Hypertension in the Elderly* (SHEP), a potential savings of six billion dollars per year can be provided by the treatment regimen of chlorthalidone with a beta blocker backup such as atenolol as compared to the combined treatment of an angiotensin converting enzyme (ACE) inhibitor with a calcium channel blocker backup. Temple (1993) also indicates that a multicooperative group study supported by health care providers is already under way to evaluate the effects of bone marrow transplant with aggressive chemotherapy for breast cancer.

1.3 REGULATORY PROCESS AND REQUIREMENTS

Chow and Liu (1995a) indicated that the development of a pharmaceutical entity is a lengthy process involving drug discovery, laboratory development, animal studies, clinical trials, and regulatory registration. The drug development can be classified into nonclinical, pre-clinical, and clinical development phases. As indicated by the *USA Today* (Feb. 3, 1993), approximately 75% of drug development is devoted to clinical development and regulatory registration. In this section we will focus on regulatory process and requirements for clinical development of a pharmaceutical entity.

For marketing approval of pharmaceutical entities, the regulatory process and requirements may vary from country (or region) to country (or region). For example, the European Community (EC), Japan, and the United States have similar but different requirements as to the conduct of clinical trials and the submission, review, and approval of clinical results for pharmaceutical entities. In this section, for simplicity, we will focus on the regulatory process and requirements for the conduct, submission, review, and approval of clinical

trials currently adopted in the United States. As was indicated earlier, the FDA was formed in 1931 to enforce the FD&C Act for marketing approval of drugs, biological products, and medical devices. With very few exceptions, since the enactment of the FD&C Act, treatment interventions such as drugs, biological products, and medical devices either currently on the market or still under investigation are the results of a joint effort between the pharmaceutical industry and the FDA. To introduce regulatory process and requirements for marketing approval of drugs, biological products, and medical devices, it is helpful to be familiar with the functional structure of the FDA.

The Food and Drug Administration

The FDA is a subcabinet organization within the Department of Health and Human Services (HHS) which is one of the major cabinets in the United States government. The FDA is headed by a commissioner with several deputy or associate commissioners to assist him or her in various issues such as regulatory affairs, management and operations, health affairs, science, legislative affairs, public affairs, planning and evaluation, and consumer affairs. Under the office of commissioner, there are currently six different *centers* of various functions for evaluation of food, drugs, and cosmetics. They are Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Center for Devices and Radiological Health (CDRH), National Center for Toxicological Research (NCTR), Center for Veterinary Medicine (CVM), and Center for Food Safety and Applied Nutrition (CFSAN).

Recently, in the interest of shortening the review process, the sponsors are required to provide the so-called user's fee for review of submission of applications to the FDA. In October 1995 CDER was reorganized to reflect the challenge of improving efficiency and shortening the review and approval process as demanded by the United States Congress and the pharmaceutical industry. Figure 1.3.1 provides the current structure of CDER at the FDA, which is composed of 10 major *offices*. These offices include Office of Management, Office of Training and Communications, Office of Compliance, Office of Information Technology, Office of Regulatory Policy, Office of Executive Program, Office of Medical Policy, Office of New Drugs, Office of Pharmaceutical Science, and Office of Pharmacoepidemiology and Statistical Science. The Office of New Drugs is responsible for drug evaluation, which consists of six offices, including Offices of Drug Evaluation I-V and Office of Pediatric Drug Development and Program Initiatives. On the other hand, Office of Pharmaceutical Science consists of four offices, including Office of New Drug Chemistry, Office of Generic Drugs, Office of Clinical Pharmacology and Biopharmaceutics, and Office of Testing and Research. Furthermore, CDER recently establishes the Office of Pharmacoepidemiology and Statistical Science in recognition of the importance of epidemiology and statistics in drug evaluation. Office of Pharmacoepidemiology and Statistical Science includes Office of Drug Safety and Office of Biostatistics. Note that each of these offices consists of several divisions. Figures 1.3.2, 1.3.3, and 1.3.4 provide respective organizations of Offices of New Drugs, Pharmaceutical Science and Pharmacoepidemiology and Statistical Science. Note that CBER has a similar functional structure though it has fewer offices than CDER.

FDA Regulations for Clinical Trials

For evaluation and marketing approval of drugs, biological products, and medical devices, the sponsors are required to submit substantial evidence of effectiveness and safety accumulated

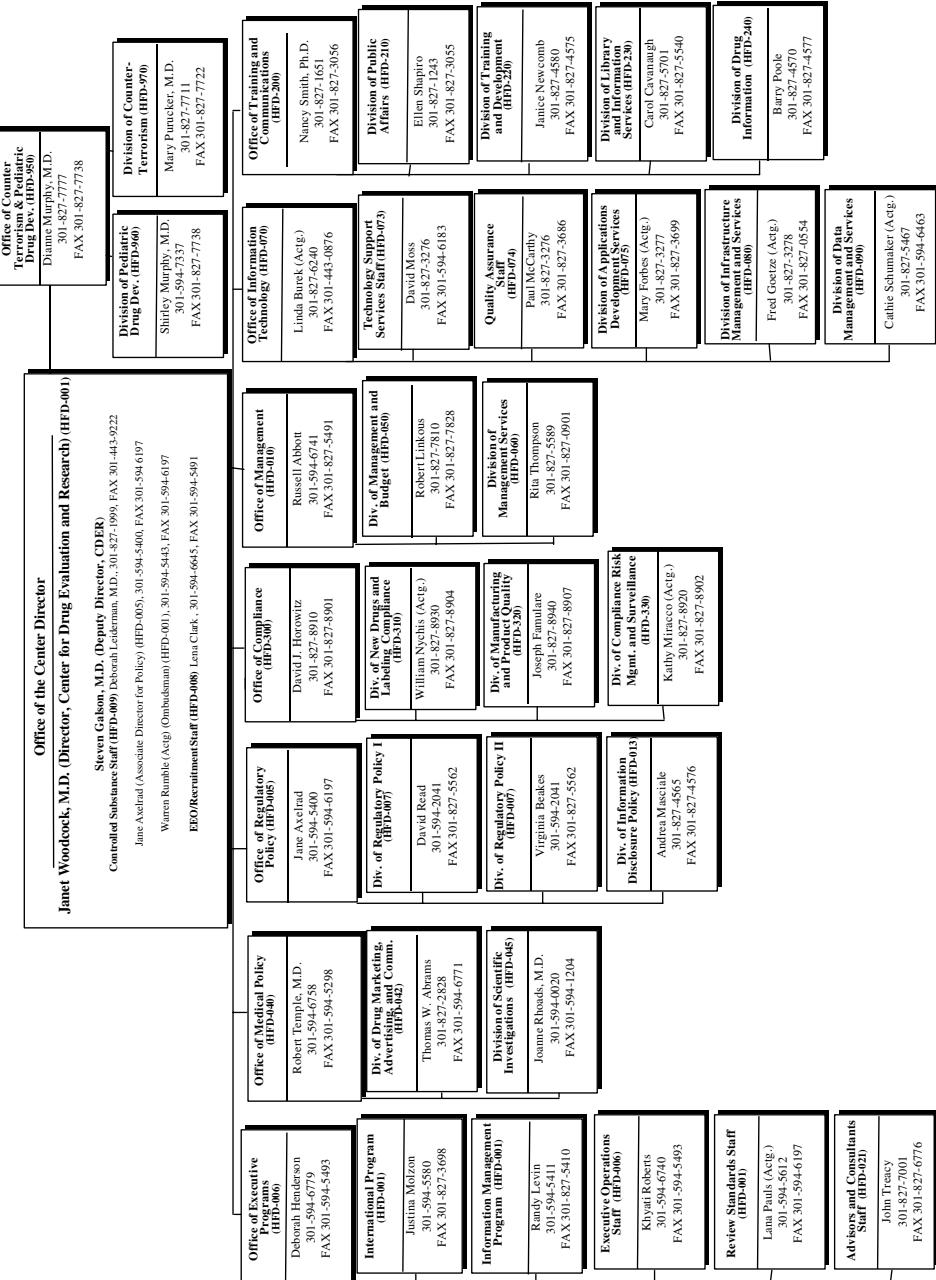


Figure 1.3.1 Center for Drug Evaluation and Research.

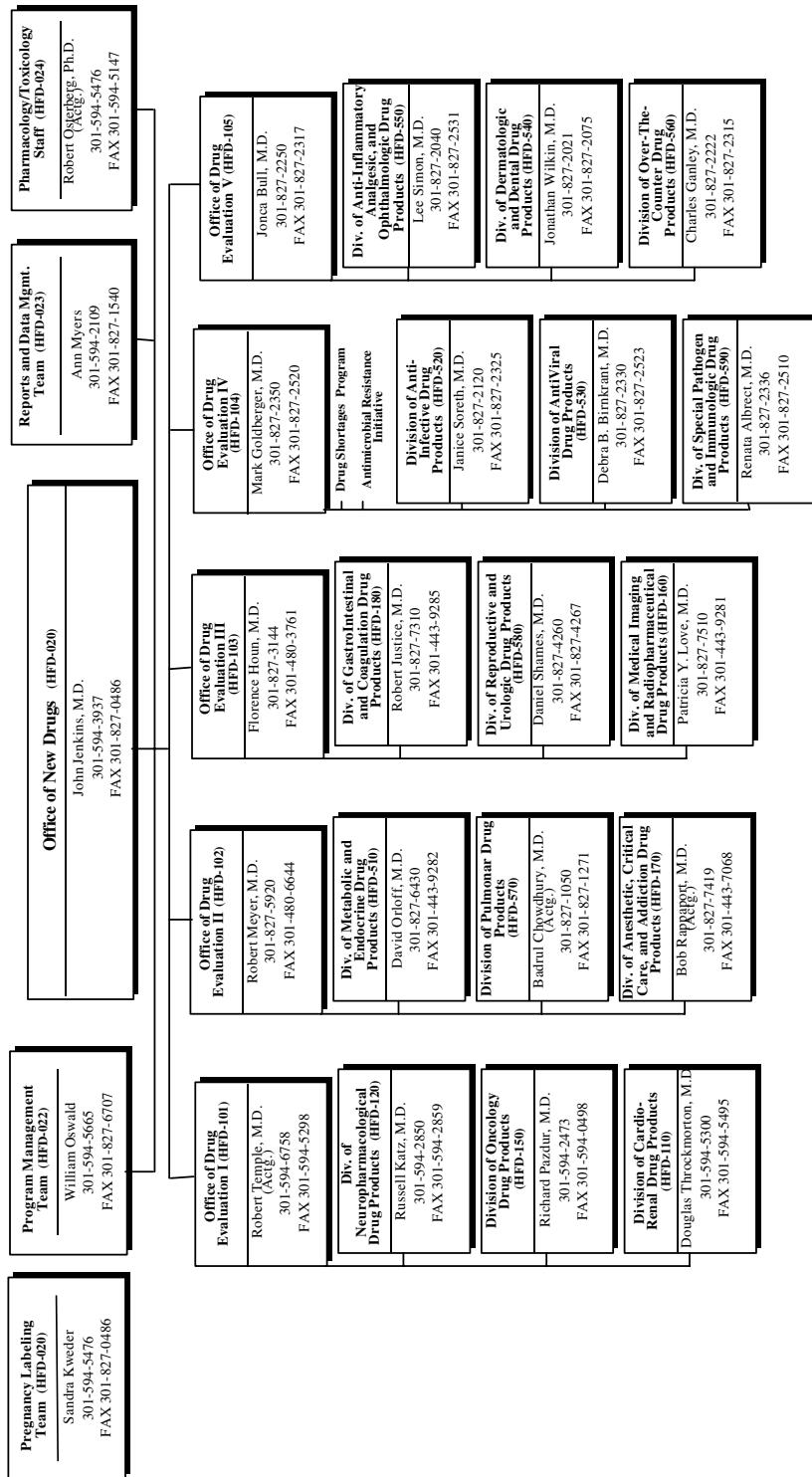


Figure 1.3.2 Office of New Drugs.

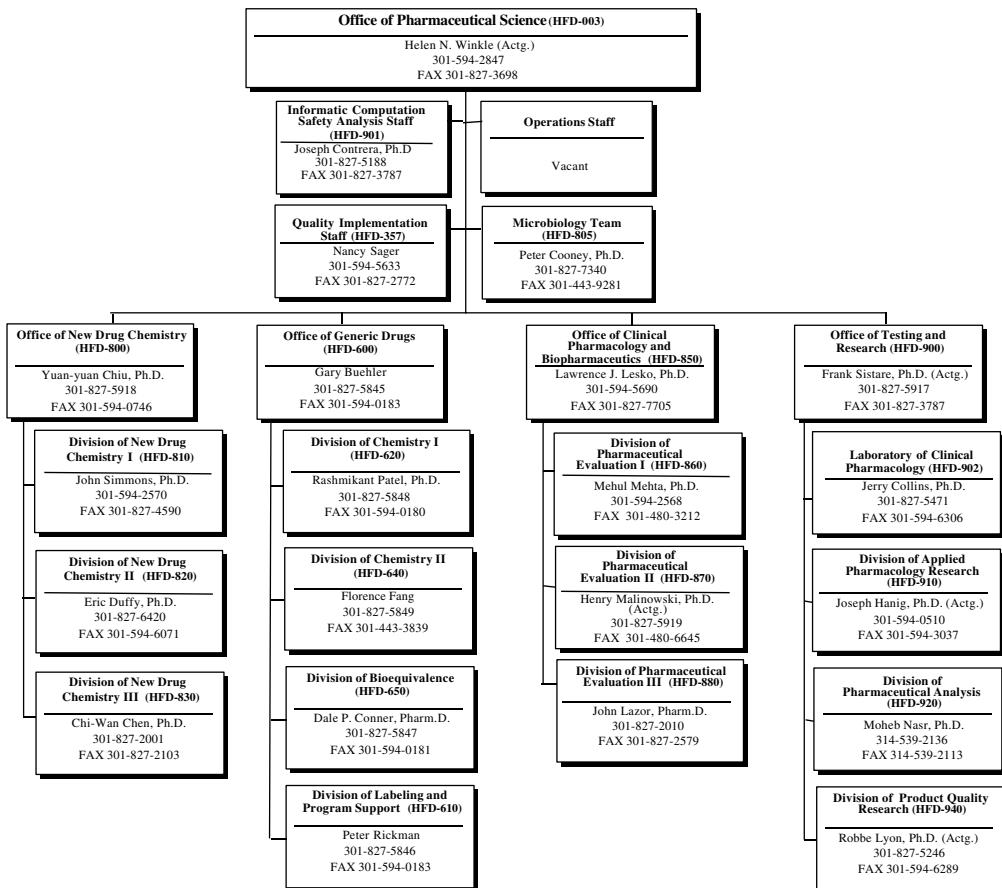


Figure 1.3.3 Office of Pharmaceutical Science.

from adequate and well-controlled clinical trials to CDER, CBER, or CDRH of the FDA, respectively. The current regulations for conducting clinical trials and the submission, review and approval of clinical results for pharmaceutical entities in the United States can be found in CFR (e.g., see 21 CFR Parts 50, 56, 312, and 314). These regulations are developed based on the FD&C Act passed in 1938. Table 1.3.1 summarizes the most relevant regulations with respect to clinical trials. These regulations cover not only pharmaceutical entities such as drugs, biological products, and medical devices under investigation but also the welfare of participating subjects and the labeling and advertising of pharmaceutical products. It can be seen from Table 1.3.1 that pharmaceutical entities can be roughly divided into three categories based on the FD&C Act and hence the CFR. These categories include drug products, biological products, and medical devices. For the first category, a drug is as defined in the FD&C Act (21 U.S.C. 321) as an article that is (1) recognized in the U.S. Pharmacopeia, official Homeopathic Pharmacopeia of the United States, or official National Formulary, or a supplement to any of them; (2) intended for use in the diagnosis, cure, mitigation, treatment, or prevention of disease in humans or other animals,

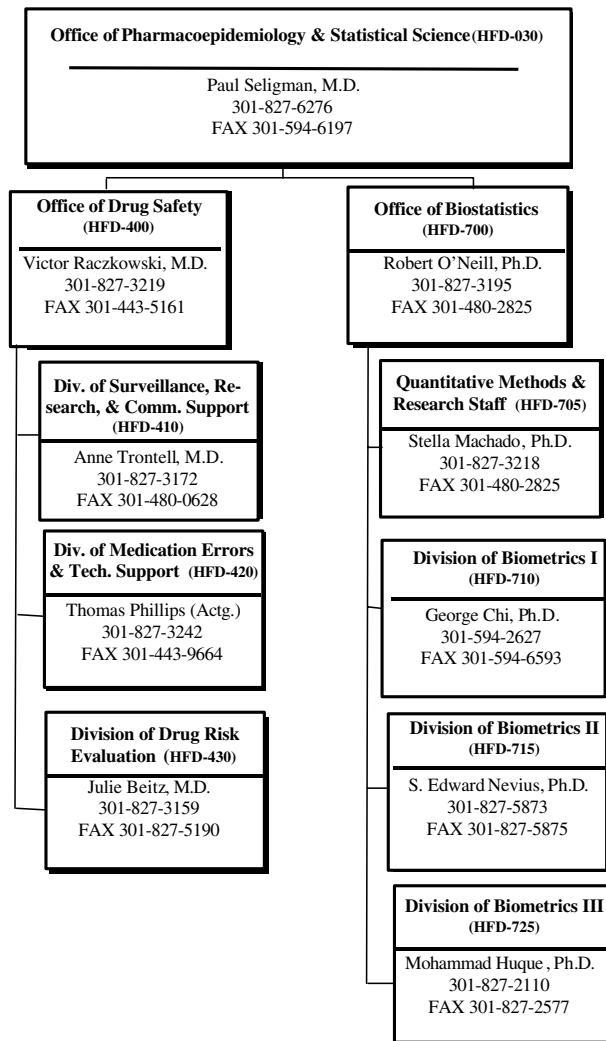


Figure 1.3.4 Office of Pharmacoepidemiology and Statistical Science.

or (3) intended to affect the structure or function of the body of humans or other animals. For the second category, a biological product is defined in the 1944 *Biologics Act* (46 U.S.C. 262) as a virus, therapeutic serum, toxin, antitoxin, bacterial or viral vaccine, blood, blood component or derivative, allergenic product, or analogous product, applicable to the prevention, treatment, or cure of disease or injuries in humans. Finally, a medical device is defined as an instrument, apparatus, implement, machine contrivance, implant, *in vitro* reagent, or other similar or related article, including any component, part, or accessory that—similar to a drug—is (1) recognized in the official National Formulary or the U.S. Pharmacopeia or any supplement in them; (2) intended for use in the diagnosis in humans or other animals; or (3) intended to affect the structure or function of the body of humans or other animals.

Table 1.3.1 U.S. Codes of Federal Regulation (CFR) for Clinical Trials Used to Approve Pharmaceutical Entities

CFR Number	Regulations
21 CFR 50	Protection of human subjects
21 CFR 56	Institutional review boards (IRB)
21 CFR 312	Investigational new drug application (IND)
Subpart E	Treatment IND
21 CFR 314	New drug application (NDA)
Subpart C	Abbreviated applications
Subpart H	Accelerated approval
21 CFR 601	Establishment license and product license applications (ELA and PLA)
Subpart E	Accelerated approval
21 CFR 316	Orphan drugs
21 CFR 320	Bioavailability and bioequivalence requirements
21 CFR 330	Over-the-counter (OTC) human drugs
21 CFR 812	Investigational device exemptions (IDE)
21 CFR 814	Premarket approval of medical devices (PMA)
21 CFR 60	Patent term restoration
21 CFR 201	Labeling
21 CFR 202	Prescription drug advertising

The CDER of the FDA has jurisdiction over administration of regulation and approval of pharmaceutical products classified as *drug*. These regulations include Investigational New Drug Application (IND) and New Drug Application (NDA) for new drugs, orphan drugs, and over-the-counter (OTC) human drugs and Abbreviated New Drug Application (ANDA) for generic drugs. On the other hand, the CBER is responsible for enforcing the regulations of biological products through processes such as an Establishment License Application (ELA) or Product License Application (PLA). Administration of the regulations for medical devices belongs to the jurisdiction of the CDRH through Investigational Device Exemptions (IDE) and Premarket Approval of Medical Devices (PMA) and other means.

A treatment for a single illness might consist of a combination of drugs, biological products, and/or medical devices. If a treatment consists of a number of drugs, then it is called a combined therapy. For example, leuprolide and flutamide are for treatment of disseminated, previously untreated D₂ stage prostate cancer. However, if a treatment consists of a combination of drugs, biologics, and/or devices such as drug with device, biologic with device, drug with biologic, drug with biologic in conjunction with device, then it is defined as a combined product. For a combined product consisting of different pharmaceutical entities, FDA requires that each of entities should be reviewed separately by appropriate centers at the FDA. In order to avoid confusion of jurisdiction over a combination product and to improve efficiency of approval process, the principle of primary mode of action of a combination product was established in the *Safe Medical Devices Act* (SMDA) in 1990 (21 U.S.C. 353). In 1992, based on this principle, three intercenter agreements were signed between CDER and CBER, between CDER and CDRH, and between CBER and

CDRH to establish the ground rules for assignment of a combined product and intercenter consultation (Margolies, 1994).

Phases of Clinical Development

In a set of new regulations promulgated in 1987 and known as the *IND Rewrite*, the phases of clinical investigation adopted by the FDA since the late 1970s is generally divided into three phases (21 CFR 312.21). These phases of clinical investigation are usually conducted sequentially but may overlap.

Phase I clinical investigation provides an initial introduction of an investigational new drug to humans. The primary objectives of phase I clinical investigation are twofold. First, it is to determine the metabolism and pharmacologic activities of the drug in humans, the side effects associated with increasing doses, and early evidence on effectiveness. In addition it is to obtain sufficient information about the drug's pharmacokinetics and pharmacological effects to permit the design of well-controlled and scientifically valid phase II clinical studies. Thus phase I clinical investigation includes studies of drug metabolism, bioavailability, dose ranging, and multiple doses. Phase I clinical investigation usually involves 20 to 80 normal volunteer subjects or patients. In general, protocols for phase I studies are less detailed and more flexible than for subsequent phases, but they must provide an outline of the investigation and also specify in detail those elements that are critical to safety. For phase I investigation, FDA's review will focus on the assessment of safety. Therefore extensive safety information such as detailed laboratory evaluations are usually collected at very intensive schedules.

Phase II studies are the first controlled clinical studies of the drug, and they involve no more than several hundred patients. The primary objectives of phase II studies are not only to initially evaluate the effectiveness of a drug based on clinical endpoints for a particular indication or indications in patients with the disease or condition under study but also to determine the dosing ranges and doses for phase III studies and the common short-term side effects and risks associated with the drug. Although the clinical investigation usually involves no more than several hundred patients, expanded phase II clinical studies may involve up to several thousand patients. Note that some pharmaceutical companies further differentiate this phase into phases IIA and IIB. Clinical studies designed to evaluate dosing are referred to as phase IIA studies, and studies designed to determine the effectiveness of the drug are called phase IIB.

Phase III studies are expanded controlled and uncontrolled trials. The primary objectives of phase III studies are not only to gather the additional information about effectiveness and safety needed to evaluate the overall benefit-risk relationship of the drug but also to provide an adequate basis for physician labeling. Phase III studies, which can involve from several hundred to several thousand patients, are performed after preliminary evidence regarding the effectiveness of the drug has been demonstrated. Note that studies performed after submission before approval are generally referred to as phase IIB studies.

In drug development, phase I studies refer to an early stage of clinical pharmacology, and phase II and III studies correspond to a later stage of clinical development. For different phases of clinical studies, the investigational processes are regulated differently, for example, the FDA review of submissions in phase I ensures that subjects are not exposed to unreasonable risks, while the review of submissions in phases II and III also ensures that the scientific design of the study is likely to produce data capable of meeting statutory standards for marketing approval.

Phase IV trials generally refer to studies performed after a drug is approved for marketing. The purpose for conducting phase IV studies is to elucidate further the incidence of adverse reactions and determine the effect of a drug on morbidity or mortality. In addition a phase IV trial is also conducted to study a patient population not previously studied such as children. In practice, phase IV studies are usually considered useful market-oriented comparison studies against competitor products.

Note that there is considerable variation within the pharmaceutical industry in categorizing clinical studies into phases. For example, in addition to phases I through IV described above, some pharmaceutical companies consider clinical studies conducted for new indications and/or new formulations (or dosage forms) as phase V studies.

1.4 INVESTIGATIONAL NEW DRUG APPLICATION

As indicated in the previous section, different regulations exist for different products, such as IND and NDA for drug products, ELA and PLA for biological products, IDE and PMA for medical devices. However, the spirit and principles for the conduct, submission, review, and approval of clinical trials are the same. Therefore, for the purpose of illustration, we will only give a detailed discussion on IND and NDA for drug products.

Before a drug can be studied in humans, its sponsor must submit an IND to the FDA. Unless notified otherwise, the sponsor may begin to investigate the drug 30 days after the FDA has received the application. The IND requirements extend throughout the period during which a drug is under study. As mentioned in Sections 312.1 and 312.3 of 21 CFR, an IND is synonymous with *Notice of Claimed Investigational Exemption for a New Drug*. Therefore an IND is, legally speaking, an exemption to the law that prevents the shipment of a new drug for interstate commerce. Consequently the drug companies that file an IND have flexibility of conducting clinical investigations of products across the United States. However, it should be noted that different states might have different laws that may require the sponsors to file separate IND to the state governments. As indicated by Kessler (1989), there are two types of INDs, commercial and noncommercial. A commercial IND permits the sponsor to gather the data on the clinical safety and effectiveness needed for an NDA. If the drug is approved by the FDA, the sponsor is allowed to market the drug for specific uses. A noncommercial IND allows the sponsor to use the drug in research or early clinical investigation to obtain advanced scientific knowledge of the drug. Note that the FDA itself does not investigate new drugs or conduct clinical trials. Pharmaceutical manufacturers, physicians, and other research organizations such as NIH may sponsor INDs. If a commercial IND proves successful, the sponsor ordinarily submits an NDA. During this period the sponsor and the FDA usually negotiate over the adequacy of the clinical data and the wording proposed for the label accompanying the drug, which sets out description, clinical pharmacology, indications and usage, contraindications, warnings, precautions, adverse reactions, and dosage and administration.

By the time an IND is filed, the sponsor should have enough information about the chemistry, manufacturing, and controls of the drug substance and drug product to ensure the identity, strength, quality, and purity of the investigational drug covered by the IND. In addition the sponsor should provide adequate information about pharmacological studies for absorption, distribution, metabolism, and excretion (ADME) and acute, subacute, and chronic toxicological studies and reproductive tests in various animal species to support that the investigational drug is reasonably safe to be evaluated in clinical trials of various durations in humans.

A very important component of an IND is the general investigational plan, which is in fact an abbreviated version of the clinical development plan for the particular pharmaceutical entity covered by the IND. However, the investigational plan should identify the phases of clinical investigation to be conducted that depend on the previous human experience with the investigational drug. Usually if a new investigational drug is developed in the United States, it is very likely that at the time of filing the IND no clinical trial on human has ever been conducted. Consequently the investigational plan might consist of all clinical trials planned for each stage of phases I, II, and III during the entire development period. On the other hand, some investigational pharmaceutical entities may be developed outside the United States. In this case sufficient human experiences may have already been accumulated. For example, for an investigational drug, suppose that the clinical development plan outside the United States has already completed phase II stage. Then the initial safety and pharmacological ADME information can be obtained from phase I clinical trials. In addition phase II dose response (ranging) studies may provide adequate dose information for the doses to be employed in the planned phase III studies. Consequently the investigational plan may only include the plan for phase III trials and some trials for specific subject population such as renal or hepatic impaired subjects. However, all information and results from phases I and II studies should be adequately documented in the section of previous human experience with the investigational drug in the IND. A general investigational plan may consist of more than one protocol depending on the stage of the clinical investigational plan to be conducted.

An IND plays an important role in the clinical development of a pharmaceutical entity. An IND should include all information about the drug product available to the company up to the time point of filing. Table 1.4.1 lists the contents of an IND provided in Section 312.23 (a) (6) of 21 CFR that a sponsor must follow and submit. A cover sheet usually refers to the form of FDA-1571. The form reinforces the sponsor's commitment to conduct the investigation in accordance with applicable regulatory requirements. A table of contents should also be included to indicate the information attached in the IND submission. The investigational plan should clearly state the rationale for the study of the drug, the indication(s) to be studied, the approach for the evaluation of the drug, the kinds of clinical trials to be conducted, the estimated number of patients, and any risks of particular severity or seriousness anticipated. For completeness, an investigator's brochure should also be provided. As mentioned earlier, the central focus of the initial IND submission should be on the general investigational plan and protocols for specific human studies. Therefore a copy of protocol(s) which includes study objectives, investigators, criteria for inclusion and exclusion, study design,

Table 1.4.1 Documents to Accompany an IND Submission

A cover sheet
A table of contents
The investigational plan
The investigator's brochure
Protocol
Chemistry, manufacturing, and controls information
Pharmacology and toxicology information
Previous human experiences with the investigational drug
Additional information
Relevant information

dosing schedule, endpoint measurements, and clinical procedure should be submitted along with the investigational plan and other information such as chemistry, manufacturing, and controls, pharmacology and toxicology, previous human experiences with the investigational drug, and any additional information relevant to the investigational drug. Note that the FDA requires that all sponsors should submit an original and two copies of all submissions to the IND file, including the original submission and all amendments and reports.

Clinical Trial Protocol

To ensure the success of an IND, a well-designed protocol is essential when conducting a clinical trial. A protocol is a plan that details how a clinical trial is to be carried out and how the data are to be collected and analyzed. It is an extremely critical and the most important document, since it ensures the quality and integrity of the clinical investigation in terms of its planning, execution, and conduct of the trial as well as the analysis of the data. Section 312.23 of 21 CFR provides minimum requirements for the protocol of a clinical trial. In addition the *Guideline for the Format and Content of the Clinical and Statistical Sections of an Application* was issued by CDER of the FDA in October 1988. Appendix C of this guideline describes key elements for a well-designed protocol. All of these requirements and elements are centered around experimental units, treatments, and evaluations of the treatments as discussed previously in Section 1.1.

Table 1.4.2 gives an example for format and contents of a well-controlled protocol for a majority of clinical trials. A well-designed protocol should always have a protocol cover sheet to provide a synopsis of the protocol. A proposed protocol cover sheet can be found in Appendix C of the FDA guideline. The objective of the study should be clearly stated at the beginning of any protocols. The study objectives are concise and precise statements of prespecified hypotheses based on clinical responses for evaluation of the drug product under study. The objectives usually consist of the primary objective, secondary objectives, and sometimes the subgroup analyses. In addition these objectives should be such that can be translated into statistical hypotheses. The subject inclusion and exclusion criteria should also be stated unambiguously in the protocol to define the targeted population to which the study results are inferred. The experimental design then employed should be able to address the study objectives with certain statistical inference. A valid experimental design should include any initial baseline or run-in periods, the treatments to be compared, the study configuration such as parallel, crossover, or forced titration, and duration of the treatment. It is extremely important to provide a description of the control groups with the rationale as to why the particular control groups are chosen for comparison.

The methods of blinding used in the study to minimize any potential known biases should be described in detail in the protocol. Likewise the protocol should provide the methods of assignments for subjects to the treatment groups. The methods of assignment are usually different randomization procedures to prevent any systematic selection bias and to ensure comparability of the treatment groups with respect to pertinent variables. Only the randomization of subjects can provide the foundation of a valid statistical inference. A well-designed protocol should describe the efficacy and safety variables to be recorded, the time that they will be evaluated, and the methods to measure them. In addition the methods for measuring the efficacy endpoints such as symptom scores for benign prostatic hyperplasia or some safety endpoints such as some important laboratory assay should be validated and results of validation need to be adequately documented in the protocol. The FDA guideline also calls for designation of primary efficacy endpoints. From

Table 1.4.2 Format and Contents of a Protocol

-
1. Protocol cover sheet
 2. Background
 3. Objectives
 - Primary
 - Secondary
 4. Study plan
 - Study design
 - Subject inclusion criteria
 - Subject exclusion criteria
 - Treatment plan
 5. Study drugs
 - Dose and route
 - Method of dispensing
 - Method and time of administration
 - Description of controls
 - Methods of randomization and blinding
 - Package and labeling
 - Duration of treatment
 - Concomitant medications
 - Concomitant procedures
 5. Measurements and observations
 - Efficacy endpoints
 - Safety endpoints
 - Validity of measurements
 - Time and events schedules
 - Screening, baseline, treatment periods, and post-treatment follow-up
 6. Statistical methods
 - Database management procedures
 - Methods to minimize bias
 - Sample size determination
 - Statistical general considerations
 - Randomization and blinding
 - Dropouts, premature termination, and missing data
 - Baseline, statistical parameters, and covariates
 - Multicenter studies
 - Multiple testing
 - Subgroup analysis
 - Interim analysis
 - Statistical analysis of demography and baseline characteristics
 - Statistical analysis of efficacy data
 - Statistical analysis of safety data
 7. Adverse events
 - Serious adverse events
 - Adverse events attributions
 - Adverse event intensity
 - Adverse event reporting
 - Laboratory test abnormalities

Table 1.4.2 (Continued)

-
- | | |
|-----|---|
| 8. | Warning and precautions |
| 9. | Subject withdrawal and discontinuation |
| | Subject withdrawal |
| | End of treatment |
| | End of study |
| 10. | Protocol changes and protocol deviations |
| | Protocol changes |
| | Protocol deviation |
| | Study termination |
| 11. | Institutional review and consent requirements |
| | Institutional review board (IRB) |
| | Informed consent |
| 12. | Obligations of investigators and |
| | administrative aspects |
| | Study drug accountability |
| | Case report forms |
| | Laboratory and other reports |
| | Study monitoring |
| | Study registry |
| | Record retention |
| | Form FDA 1572 |
| | Signatures of investigators |
| | Confidentiality |
| | Publication of results |
| 13. | Flow chart of studies activities |
| 14. | References |
| 15. | Appendices |
-

the primary objective based on the primary efficacy endpoint, the statistical hypothesis for sample size determination can be formulated and stated in the protocol. The treatment effects assumed in both null and alternative hypotheses with respect to the experimental design employed in the protocol and the variability assumed for sample size determination should be described in full detail in the protocol as should the procedures for accurate, consistent, and reliable data. The statistical method section of any protocols should address general statistical issues often encountered in the study. These issues include randomization and blinding, handling of dropouts, premature termination of subjects, and missing data, defining the baseline and calculation of statistical parameters such as percent change from baseline and use of covariates such as age or gender in the analysis, the issues of multicenter studies, and multiple comparisons and subgroup analysis.

If interim analyses or administrative looks are expected, the protocol needs to describe any planned interim analyses or administrative looks of the data and the composition, function, and responsibilities of a possible outside data-monitoring committee. The description of interim analyses consists of monitoring procedures, the variables to be analyzed, the frequency of the interim analyses, adjustment of nominal level of significance, and decision rules for termination of the study. In addition the statistical methods for analyses of demography and baseline characteristics together with the various efficacy and safety endpoints should be described fully in the protocol. The protocol must define adverse events, serious adverse events, and attributions and intensity of adverse events and describe how

the adverse events are reported. Other ethical and administration issues should also be addressed in the protocol. They are warnings and precautions, subject withdrawal and discontinuation, protocol changes and deviations, institutional review board and consent form, obligation of investigators, case report form, and others.

It should be noted that once an IND is in effect, the sponsor is required to submit a protocol amendment if there are any changes in protocol that significantly affect the subjects' safety. Under 21 CFR 312.30(b) several examples of changes requiring an amendment are given. These examples include (1) any increase in drug dosage, duration, and number of subjects, (2) any significant change in the study design, (3) the addition of a new test or procedure that is intended for monitoring side effects or an adverse event. In addition the FDA also requires an amendment be submitted if the sponsor intends to conduct a study that is not covered by the protocol. As stated in 21 CFR 312.30(a) the sponsor may begin such study provided that a new protocol is submitted to the FDA for review and is approved by the institutional review board. Furthermore, when a new investigator is added to the study, the sponsor must submit a protocol amendment and notify FDA of the new investigator within 30 days of the investigator being added. Note that modifications of the design for phase I studies that do not affect critical safety assessment are required to be reported to FDA only in the annual report.

Institutional Review Board

Since 1971 the FDA has required that all proposed clinical studies be reviewed both by the FDA and an institutional review board (IRB). The responsibility of an IRB is not only to evaluate the ethical acceptability of the proposed clinical research but also to examine the scientific validity of the study to the extent needed to be confident that the study does not expose its subjects to unreasonable risk (Petrcciani, 1981). This IRB is formally designated by a public or private institution in which research is conducted to review, approve, and monitor research involving human subjects. Each participating clinical investigator is required to submit all protocols to an IRB. An IRB must formally grant approval before an investigation may proceed, which is in contrast to the 30-day notification that the sponsors must give the FDA. To ensure that the investigators are included in the review process, the FDA requires that the clinical investigators communicate with the IRB. The IRB must monitor activities within their institutions.

The composition and function of an IRB are subject to FDA requirements. Section 56.107 in Part 56 of 21 CFR states that each IRB should have at least five members with varying backgrounds to promote a complete review of research activities commonly conducted by the institution. In order to avoid conflict of interest and to provide an unbiased and objective evaluation of scientific merits, ethical conduct of clinical trials, and protection of human subjects, the CFR enforces a very strict requirement for the composition of members of an IRB. The research institution should make every effort to ensure that no IRB is entirely composed of one gender. In addition no IRB may consist entirely of members of one profession. In particular, each IRB should include at least one member whose primary concerns are in the scientific area and at least one member whose primary concerns are in nonscientific areas. On the other hand, each IRB should include at least one member who is not affiliated with the institution and who is not part of the immediate family of a person who is affiliated with the institution. Furthermore no IRB should have a member participate in the IRB's initial or continuous review of any project in which the member has a conflicting interest, except to provide information requested by the IRB.

Safety Report

The sponsor of an IND is required to notify FDA and all participating investigators in a written IND safety report of any adverse experience associated with use of the drug. Adverse experiences need to be reported include serious and unexpected adverse experiences. A serious adverse experience is defined as any experience that is fatal, life-threatening, requiring inpatient hospitalization, prolongation of existing hospitalization, resulting in persistent or significant disability/incapacity, or congenital anomaly/birth defect. An unexpected adverse experience is referred to as any adverse experience that is not identified in nature, severity, or frequency in the current investigator brochure or the general investigational plan or elsewhere in the current application, as amended.

The FDA requires that any serious and unexpected adverse experience associated with use of the drug in the clinical studies conducted under the IND be reported in writing to the agency and all participating investigators within 10 working days. The sponsor is required to fill out the FDA-1639 form to report an adverse experience. Fatal or immediately life-threatening experience require a telephone report to the agency within three working days after receipt of the information. A follow-up of the investigation of all safety information is also expected.

Treatment IND

During the clinical investigation of the drug under an IND, it may be necessary and ethical to make the drug available to those patients who are not in the clinical trials. Since 1987 the FDA permits an investigational drug to be used under a treatment protocol or treatment IND if the drug is intended to treat a serious or immediately life-threatening disease, especially when there is no comparable or satisfactory alternative drug or other therapy available to treat that stage of the disease in the intended patient population. FDA, however, may deny a request for treatment use of an investigational drug under a treatment protocol or treatment IND if the sponsor fails to show that the drug may be effective for its intended use in its intended patient population or that the drug may expose the patients to an unreasonable and significant additional risk of illness or injury.

Withdraw and Termination of an IND

At any time a sponsor may withdraw an effective IND without prejudice. However, if an IND is withdrawn, FDA must be notified and all clinical investigations conducted under the IND shall be ended. If an IND is withdrawn because of a safety reason, the sponsor has to promptly inform FDA, all investigators, and all reviewing IRBs with the reasons for such withdrawal.

If there are any deficiencies in the IND or in the conduct of an investigation under an IND, the FDA may terminate an IND. If an IND is terminated, the sponsor must end all clinical investigations conducted under the IND and recall or dispose all unused supplies of the drug. Some examples of deficiencies in an IND are discussed under 21 CFR 312.44. For example, FDA may propose to terminate IND if it finds that human subjects would be exposed to an unreasonable and significant risk of illness or injury. In such a case the FDA will notify the sponsor in writing and invite correction or explanation within a period of 30 days. A terminated IND is subject to reinstatement based on additional submissions that eliminate such risk. In this case a regulatory hearing on the question of whether the IND should be reinstated will be held.

Communication with the FDA

FDA encourages open communication regarding any scientific or medical question that may be raised during the clinical investigation. Basically it is suggested that such communication be arranged at the end of the phase II study and prior to a marketing application. The purpose of an end-of-phase II meeting is to review the safety of the drug proceeding to phase III. This meeting is helpful not only in that it evaluates the phase III plan and protocols but also in that it identifies any additional information necessary to support a marketing application for the uses under investigation. Note that a similar meeting may be held at the end of phase I in order to review results of tolerance/safety studies and the adequacy of the remaining development program. At the end of phase I, a meeting would be requested by a sponsor when the drug or biologic product is being developed for a life-threatening disease and the sponsor wishes to file under the expedited registration regulations. The purpose of pre-NDA meetings is not only to uncover any major unresolved problems but also to identify those studies that are needed for establishment of drug effectiveness. In addition the communication enables the sponsor to acquaint FDA reviewers with the general information to be submitted in the marketing application. More important, the communication provides the opportunity to discuss (1) appropriate methods for statistical analysis of the data and (2) the best approach to the presentation and formatting of the data.

1.5 NEW DRUG APPLICATION

For approval of a new drug, the FDA requires at least two adequate well-controlled clinical studies be conducted in humans to demonstrate substantial evidence of the effectiveness and safety of the drug. The *substantial evidence* as required in the Kefauver-Harris amendments to the FD&C Act in 1962 is defined as the evidence consisting of adequate and well-controlled investigations, including clinical investigations, by experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved, on the basis of which it could fairly and responsibly be concluded by such experts that the drug will have the effect it purports to represent under the conditions of use prescribed, recommended, or suggested in the labeling or proposed labeling thereof. Based on this amendment, the FDA requests that reports of adequate and well-controlled investigations provide the primary basis for determining whether there is *substantial evidence* to support the claims of new drugs and antibiotics. Section 314.126 of 21 CFR provides the definition of an adequate and well-controlled study, which is summarized in Table 1.5.1. It can be seen from Table 1.5.1 that an adequate and well-controlled study is judged by eight criteria specified in the CFR. These criteria are objectives, method of analysis, design of studies, selection of subjects, assignment of subjects, participants of studies, assessment of responses, and effect. First, each study should have a very clear statement of objectives for clinical investigation such that they can be reformulated into statistical hypotheses and estimation procedures. In addition proposed methods of analyses should be described in the protocol and actual statistical methods used for analyses of data should be described in detail in the report. Second, each clinical study should employ a design that allows a valid comparison with a control for an unbiased assessment of drug effect. Therefore selection of a suitable control is one of keys to integrity and quality of an adequate and well-controlled study. The CFR recognizes the following controls: placebo concurrent control, dose-comparison concurrent control, no treatment control, active concurrent control, and historical control. Next, the subjects in the study should have the disease or condition

Table 1.5.1 Characteristics of an Adequate and Well-Controlled Study

Criteria	Characteristics
Objectives	Clear statement of investigation's purpose
Methods of analysis	Summary of proposed or actual methods of analysis
Design	Valid comparison with a control to provide a quantitative assessment of drug effect
Selection of subjects	Adequate assurance of the disease or conditions under study
Assignment of subjects	Minimization of bias and assurance of comparability of groups
Participants of studies	Minimization of bias on the part of subjects, observers, and analysts
Assessment of responses	Well-defined and reliable
Assessment of the effect	Requirements of appropriate statistical methods

under study. Furthermore subjects should be randomly assigned to different groups in the study to minimize potential bias and ensure comparability of the groups with respect to pertinent variables such as age, gender, race, and other important prognostic factors. All statistical inferences are based on such randomization and possibly stratification to achieve these goals. However, bias will still occur if no adequate measures are taken on the part of subjects, investigator, and analysts of the study. Therefore blinding is extremely crucial to eliminate the potential bias from this source. Usually an adequate and well-controlled study is at least double blinded whereby investigators and subjects are blinded to the treatments during the study. However, currently a triple-blind study in which the sponsor (i.e., clinical monitor) of the study is also blinded to the treatment is not uncommon. Another critical criterion is the validity and reliability of assessment of responses. For example, the methods for measurement of responses such as symptom scores for benign prostate hyperplasia should be validated before their usage in the study (Barry et al., 1992). Finally, appropriate statistical methods should be used for assessment of comparability among treatment groups with respect to pertinent variables mentioned above and for unbiased evaluation of drug effects.

Section 314.50 of 21 CFR specifies the format and content of an NDA, which is summarized in Table 1.5.2. The FDA requests that the applicant should submit a complete archival copy of the new drug application form (A) to (F) with a cover letter. In addition, the sponsor needs to submit a review copy for each of the six technical sections with the cover letter, application form (356H) of (A), index of (B), and summary of (C) as given in Table 1.5.2 to each of six reviewing disciplines. The reviewing disciplines include chemistry reviewers for the chemistry, manufacturing, and controls; pharmacology reviewers for nonclinical pharmacology and toxicology; medical reviewers for clinical data section; and statisticians for statistical technical section. The outline of review copies for clinical reviewing divisions include (1) cover letter, (2) application form (356H), (3) index, (4) summary, and (5) clinical section. The outline of review copies for statistical reviewing division consists of (1) cover letter, (2) application form (356H), (3) index, (4) summary, and (5) statistical section.

Table 1.5.3 provides a summary of the format and content of a registration dossier for the European Economic Community (EEC). A comparison of Table 1.5.2 and Table 1.5.3 reveals that the information required by the FDA and ECC for marketing approval of a drug is essentially the same. However, no statistical technical section is required in the ECC

Table 1.5.2 A Summary of Contents and Format of a New Drug Application (NDA)

Cover letter
A. Application form (365H)
B. Index
C. Summary
D. Technical sections
1. Chemistry, manufacturing, and controls
2. Nonclinical pharmacology and toxicology
3. Human pharmacology and bioavailability
4. Microbiology (for anti-infective drugs)
5. Clinical data
6. Statistical
E. Samples and labeling
F. Case report forms and tabulations
1. Case report tabulations
2. Case report forms
3. Additional data

Note: Based on Section 314.50 of Part 21 of Codes of Federal Regulation (4-1-94 edition).

registration. In October 1988, to assist an applicant in presenting the clinical and statistical data required as part of an NDA submission, the CDER of the FDA issued the *Guideline for the Format and Content of the Clinical and Statistical Sections of an Application* under 21 CFR 314.50, which is summarized in Table 1.5.4. The guideline indicates the preference of having one integrated clinical and statistical report rather than two separate reports. A complete submission should include clinical section [21 CFR 314.50(d)(5)], statistical section [21 CFR 314.50(d)(6)], and case report forms and tabulations [21 CFR 314.50(f)]. The same guideline also provides the content and format of the fully integrated clinical and statistical report of a controlled clinical study in an NDA. A summary of it is given in Table 1.5.5. Based on the content and format of the fully integrated and statistical report of a controlled study required by the FDA, the *Structure and Content of Clinical Study Reports* was also issued by the European Community in May 1993. A summary is given in Table 1.5.6. In addition the European Community also published a guideline entitled *Bio-statistical Methodology in Clinical Trials in Applications for Marketing Authorizations for Medicinal Products* in March 1993.

Expanded Access

A standard clinical development program of phases I, II, and III clinical trials and traditional approval of a new pharmaceutical entity through IND and NDA processes by the FDA will generally take between 8 to 12 years with an average cost around \$500 million. Kessler and Feiden (1995) indicated that on average, the FDA receives around 100 original NDAs each year. For each NDA submission, FDA requires substantial evidence of efficacy and safety be provided with fully matured and complete data generated from at least two adequate and well-controlled studies before it can be considered for approval. This requirement is necessary for drugs with marginal clinical advantages and for

Table 1.5.3 Format and Contents of a Registration Dossier for the European Economic Community (EEC)

Flyleaf	
Annex I:	General information
Annex II:	Information and documents on physicochemical, biological, or microbiological tests
Annex II.A:	Complete qualitative and quantitative composition
Annex II.B:	Method of preparation
Annex II.C:	Controls of starting materials
Annex II.D:	Control tests on intermediate products (if necessary)
Annex II.E:	Control tests for the finished product
Annex II.F:	Stability tests
Annex II.G:	Conclusions
Annex III:	Toxicological and pharmacological tests
Annex III.A:	Acute toxicity
Annex III.B:	Toxicity with repeated administration
Annex III.C:	Fetal toxicity
Annex III.D:	Fertility studies
Annex III.E:	Carcinogenicity and mutagenicity
Annex III.F:	Pharmacodynamics
Annex III.G:	Pharmacokinetics
Annex IV:	Clinical trials
Annex IV.A:	Human pharmacology
Annex IV.B:	Clinical data
Annex IV.C:	Side effects and interactions
Annex V:	Special particulars
Annex V.A:	Dosage forms
Annex V.B:	Samples
Annex V.C:	Manufacturing authorization
Annex V.D:	Marketing authorization

Table 1.5.4 Summary of the Clinical and Statistical Section of an NDA

- A. List of investigators; list of INDs and NDAs
- B. Background/overview of clinical investigations
- C. Clinical pharmacology
- D. Control clinical studies
- E. Uncontrolled clinical studies
- F. Other studies and information
- G. Integrated summary of effectiveness data
- H. Integrated summary of safety data
- I. Drug abuse and overdosage
- J. Integrated summary of benefits and risks of the drug

Source: Based on *Guideline for the Format and Content of the Clinical and Statistical Sections of an Application* (July, 1988, Center for Drug Evaluation and Research, FDA).

Table 1.5.5 Summary of Format and Contents of a Fully Integrated Clinical and Statistical Report for a Controlled Study in an NDA

-
- A. Introduction
 - B. Fully integrated clinical and statistical report of a controlled clinical study
 - 1. Title page
 - 2. Table of contents for the study
 - 3. Identity of the test materials, lot numbers, etc.
 - 4. Introduction
 - 5. Study objectives
 - 6. Investigational plan
 - 7. Statistical methods planned in the protocol
 - 8. Disposition of patients entered
 - 9. Effectiveness results
 - 10. Safety results
 - 11. Summary and conclusion
 - 12. References
 - 13. Appendices
-

Source: Based on *Guideline for the Format and Content of the Clinical and Statistical Sections of an Application* (July, 1988, Center for Drug Evaluation and Research, FDA).

treatment of conditions or diseases that are not life-threatening. However, if the diseases are life-threatening or severely debilitating, then the traditional clinical development and approval process might not be soon enough for the subjects whose life may be saved by the promising drugs. According to Section 312.81 in 21 CFR, life-threatening diseases are defined as (1) the diseases or conditions where the likelihood of death is high unless the course of the disease is interrupted and (2) diseases or conditions with potentially fatal outcomes, where the endpoint of clinical trial analysis is survival. On the other hand,

Table 1.5.6 Summary of Format and Contents of Clinical Study Reports for the European Economic Community (EEC)

-
- 1. Title page
 - 2. Table of contents for the study
 - 3. Synopsis
 - 4. Investigators
 - 5. Introduction
 - 6. Study objectives
 - 7. Investigational plan
 - 8. Study subjects
 - 9. Effectiveness evaluation
 - 10. Safety evaluation
 - 11. Discussion
 - 12. Overall conclusions
 - 13. Summary tables, figures, and graphs cited in text
 - 14. Reference list
 - 15. Appendices
-

Source: Based on Structure and Content of Clinical Study Reports Joint EFPIA/CPMP Document—May 13, 1993.

severely debilitating diseases are those that cause major irreversible morbidity. Since 1987 regulations have been established for early access to promising experimental drugs and for accelerated approval of drugs for treatment of life-threatening or severely debilitating diseases.

Expanded access is devised through treatment IND (Section 312.34 of 21 CFR) and parallel track regulations. For a serious or immediately life-threatening disease with no satisfactory therapy available, as mentioned before, a treatment IND allows promising new drugs to be widely distributed even when data and experience are not sufficient enough for a full marketing approval. On the other hand, for example, for the patients infected with human immunodeficiency virus (HIV) who are not qualified for clinical trials and have no other alternative treatment, parallel track regulations issued in 1992 provide a means for these patients to obtain experimental therapy very early in the development stage through their private physicians. In 1992 the FDA also established the regulations for accelerated approval of the drug for serious or life-threatening diseases based on a surrogate clinical endpoint other than survival or irreversible morbidity (Subpart H of Section 314 in 21 CFR). A new concept for approval called *Telescoping Trials* has also emerged (Kessler and Feiden, 1995). Under this concept, phase III clinical trials might be totally eliminated. For example, the FDA might consider approval of a drug for a serious disease which, during phase II clinical trials, demonstrates a positive impact on survival or irreversible morbidity. The time table for drug evaluation and approval is illustrated in Figure 1.5.1 which is adopted from Kessler and Feiden (1995). A successful example of expanded access and accelerated approval provided by these regulations is the review and approval of dideoxyinosine (ddI) of Bristol-Myers Squibb Company for patients with HIV. An expanded access to ddI was initiated in September 1989. The new drug application based on the data of phase I clinical trials with no control group was filed in April 1991. The FDA granted conditional approval of the drug in October 1991 based on a clinical surrogate end point called a CD4+ lymphocyte count. With the data from phases II and III clinical trials submitted in April 1992, the approval of ddI was broadened in September 1992. The history of ddI case is illustrated in Figure 1.5.2 (also adopted from Kessler and Feiden, 1995). Another example for fast-track development and accelerated approval is the case of fludarabine phosphate (fludara) for treatment of refractory chronic lymphocytic leukemia (CLL) (Tessman, Gipson, and Levins, 1994). Fludara is the first new drug approved for this common form of adult leukemia in the United States over 50 years. The NDA, filed in November 1989 and approved in April 1991, was in fact based on retrospective analyses of phase II clinical trials conducted by NCI through cooperative groups including Southwest Oncology Group (SWOG) and M.D. Anderson Cancer Center in Houston, Texas. In addition an early excess to the drug was provided in 1989 through NCI's Group C protocol, which is equivalent to NCI's version of treatment IND. The last example is the approval of Gleevec (omatinib mesylate) for oral treatment for patients with chronic myeloid leukemia (CML) by the U.S. FDA in 2001. Gleevec is a specific inhibitor of tyrosine kinase enzymes that plays an important role in CML. Under accelerated approval regulation and orphan drug status, the U.S. FDA reviewed and approved the marketing application in less than 3 months. This approval for three phases of CML was based on separate single-arm studies using surrogate endpoints such as major cytogenetic response. One of these studies was recently published (Kantarjian et al., 2002). However, the then-U.S. FDA acting commissioner, B. A. Schwetz, D.V.M., Ph.D., indicated that further studies are needed to evaluate whether Gleevec provides an actual clinical benefit, such as improved survival.

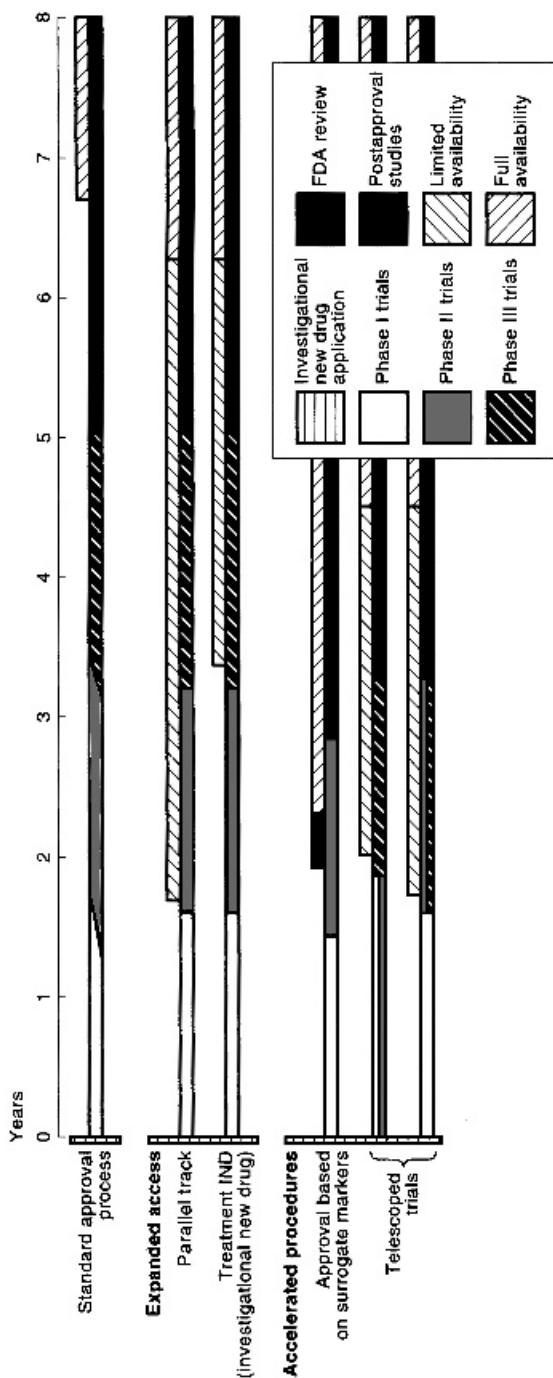


Figure 1.5.1 Timeline for drug evaluation and approval. Drug development can take many years of successively larger clinical studies (*top*). The FDA's expanded-access rules now make available to patients serious illness drugs that are still under investigation (*middle*). The agency has also reduced the time it takes to approve new drugs for sale, and it may issue provisional approval for widespread marketing of a compound on the basis of significantly fewer data than it once required (*bottom*). A drug may then be removed from the market if later evaluations show that it is not beneficial. (Source: Kessler and Faiden, 1995.)

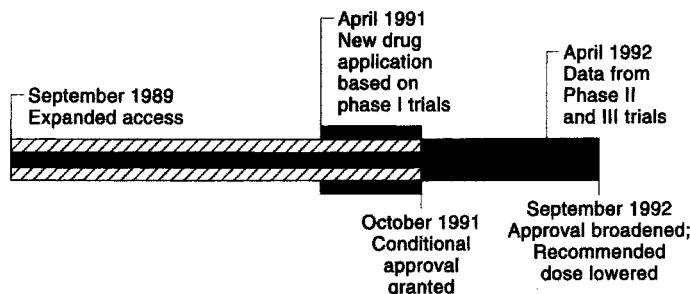


Figure 1.5.2 Case history of dideoxyinosine (ddI) shows how efforts to streamline the regulatory process have paid off. In September 1989 the drug was made available to many AIDS patients on an expanded-access basis. The FDA approved the drug for sale after reviewing preliminary results from ongoing studies and then expanded the approval once final results came in. (Source: Kessler and Faiden, 1995.)

Abbreviated New Drug Application

An abbreviated NDA (ANDA) is usually reserved for drug products (e.g., generics) that duplicate products previously approved under a full NDA. For an ANDA, reports of non-clinical laboratory studies and clinical investigations except for those pertaining to *in vivo* bioavailability of the drug product are not required. The information may be omitted when the FDA has determined that the information already available to it is adequate to establish that a particular dosage form of a drug meets the statutory standards for safety and effectiveness. The duplicate products are usually referred to as products with the same active ingredient(s), route of administration, dosage form, strength, or condition of use that may be made by different manufacturers.

As mentioned earlier, under the *Drug Price Competition and Patient Term Restoration Act* passed in 1984, the FDA may approve generic drug products if the generic drug companies can provide evidence that the rates and extents of absorption of their drug products do not show a significant difference from those of the innovator drug products when administered at the same molar dose of the therapeutic moiety under similar experimental conditions (21 CFR 320). The *Drug Price Competition and Patent Term Restoration Act* states FDA's authority for all generic drug approvals through an ANDA submission for bioequivalence review. An ANDA submission should include product information, pharmacokinetic data and analysis, statistical analysis, analytical methodology and validation, and clinical data. In the ANDA submission the FDA requires the sponsor to provide necessary information regarding the drug product such as formulation, potency, expiration dating period (or shelf life), and dissolution data. For example, the dissolution profile of the generic drug product should be comparable with that of the innovator drug product for drug release. Before the conduct of a bioavailability and bioequivalence study, the FDA also requires the sponsor to provide validation data for the analytical method used in the study. The analytic method should be validated according to standards specified in the U.S. Pharmacopeia and National Formulary (USP/NF, 2002). For example, the analytical method needs to be validated in terms of its accuracy, precision, selectivity, limit of detection, limit of quantitation, range, linearity, and ruggedness (Chow and Liu, 1995a). For pharmacokinetic data, descriptive statistics should be given by the sampling time point and for each pharmacokinetic responses. To ensure the validity of bioequivalence assessment,

the Division of Bioequivalence, Office of Generic Drugs of CDER at the FDA issued a *Guidances on Bioavailability and Bioequivalence Studies for Orally Administrated Drug Products—General Consideration* in March, 2003 and *Statistical Approaches to Establishing Bioequivalence* in January 2001, respectively. The guidance sets forth regulations for valid statistical analysis for bioequivalence assessment. Note that detailed information regarding statistical design and analysis of bioavailability and bioequivalence studies can be found in Chow and Liu (2000). In addition any relevant clinical findings, adverse reactions and deviation from the protocol need to be included in the ANDA submission.

Supplemental New Drug Application

A supplemental NDA (SNDA) is referred to as documentation submitted to FDA on a drug substance or product that is already the subject of an approved NDA. Supplements may be submitted for a variety of reasons such as labeling changes, a new or expanded clinical indication, or a new dosage form. For example, for labeling changes, the sponsor may want to add a new specification or test method or changes in the methods, facility, or controls to provide increased assurance that the drug will have the characteristics of identity, strength, quality, and purity that it purports to possess. For drug substance and/or drug product, the sponsor may want to relax the limits for a specification, establish a new regulatory analytical method, or delete a specification or regulatory analytical method. In addition the sponsor may want to extend the expiration date of the drug product based on data obtained under a new or revised stability testing protocol that has not been approved in the application or to establish a new procedure for reprocessing a batch of the drug product that fails to meet specification. It, however, should be noted that in an SNDA, the sponsor is required to fully describe the change in each condition established in an approved application beyond the variation already provided for in the application.

Advisory Committee

The FDA has established advisory committees each consisting of clinical, pharmacological, and statistical experts and one consumer advocate (not employed by the FDA) in designated drug classes and subspecialties. The responsibilities of the committees are to review data presented in NDA's and to advise FDA as to whether there exists substantial evidence of safety and effectiveness based on adequate and well-controlled clinical studies. In addition the committee may also be asked at times to review certain INDs, protocols, or important issues relating to marketed drugs and biologics. The advisory committees not only supplement the FDA's expertise but also allow an independent peer review during the regulatory process. Note that the FDA usually prepares a set of questions for the advisory committee to address at the meeting. The following is a list of some typical questions:

- 1. Are there two or more adequate and well-controlled trials?**
- 2. Have the patient populations been well enough characterized?**
- 3. Has the dose-response relationship been sufficiently characterized?**
- 4. Do you recommend the use of the drug for the indication sought by the sponsor for the intended patient population?**

The FDA usually will follow the recommendations made by the Advisory Committee for marketing approval, though they do not have to legally.

1.6 CLINICAL DEVELOPMENT AND PRACTICE

Clinical research and development in pharmaceutical environment is to scientifically evaluate the benefits and risks of promising pharmaceutical entities at a minimal cost and within a very short timeframe. To ensure the success of the development of the pharmaceutical entity, a clinical development plan is necessary.

Clinical Development Plan

A clinical development plan (CDP) is a description of clinical studies that will be carried out in order to assess the safety and effectiveness of the drug. A clinical development plan typically includes a development rationale, listing of trial characteristics, timeline, cost, and resource requirements. A good and flexible clinical development plan hence is extremely crucial and important to the success and unbiased assessment of a potential pharmaceutical entity. Although a typical CDP is based primarily on the validity of medical and scientific considerations, other factors that involve issues such as biostatistics, regulatory, marketing, and management are equally important. For a successful CDP, we first need to define a product profile for the promising pharmaceutical entity before any clinical development. Table 1.6.1 lists essential components of a product profile. These components set the goals and objectives for the clinical development program of a pharmaceutical entity. A clinical development program is referred to as the set of different clinical trial plans at different stages with milestones for assessment and decision making to evaluate the goals and objectives stated in the product profile. For example, if the drug product under development is for an indication intended for a particular population, the relative merits and disadvantages of the product as compared to other products either on the market or still under development should objectively be assessed. In order to evaluate the relative merits, minimum requirements and termination criteria on the effectiveness and safety of the product are usually set. These requirements and criteria are evaluated through statistical analysis of data collected from a series of clinical trials. The deadlines for milestones and decision making should also be scheduled in CDP according to the time when certain clinical trials to evaluate the requirements and criteria are completed and the data are adequately analyzed. Since a huge investment is usually necessarily committed to develop a new pharmaceutical entity, information based on efficacy and safety alone may not be enough to evaluate a potential product. It is therefore recommended that cost-effectiveness and quality of life be evaluated, especially for the me-too products in a saturated market. In this case requirements and criteria for cost-effectiveness and quality of life need to be included at milestones and/or decision-making points. As indicated earlier, although many factors such as statistics, marketing, regulatory, and management need to be considered in a CDP, the scientific validity of clinical investigations is the key to the success of a clinical development program.

In the pharmaceutical industry clinical development of a pharmaceutical entity starts with seeking alternatives or new drug therapies for an existing health problem (e.g., hypertension) or a newly identified health problem (e.g., AIDS). The health problem of interest may be related to virus, cardiovascular, cancer diseases, or other diseases. Once the health problem is selected or identified, whether it is worth developing an alternative or a new pharmaceutical entity for this particular disease is a critical development decision point. A clear decision point can increase the success of the project and consequently reduce the risk and cost. Suppose that it is decided to proceed with the development of a pharmaceutical

Table 1.6.1 Components of a Pharmaceutical Product Profile

Target population
Innovation potentials
Therapeutic concepts
Innovative elements
Technological advances
Patent status
Route of administrations
Doses
Formulations
Regimens
Duration of dosing
Status of market
Current competitors
On market
Under development
Advantages
Disadvantages
Minimum requirements
Efficacy
Safety
Termination criteria
Efficacy
Safety
Time frames
Milestones

entity (e.g., enzymes or receptors), a number of chemical modifications and ADME tests in animals may be necessary before it can be tested on humans. ADME studies are used to determine how a drug is taken up by the body, where it goes in the body, the chemical changes it undergoes in the body, and how it is eliminated from the body. ADME studies describe the pharmacokinetics and bioavailability of a drug. If the drug shows promising effectiveness and safety in animals, the sponsor normally will make a decision to go for an IND. As indicated in Section 1.4, an IND is a synthesis process that includes formulation, analytical method development and validation, stability, animal toxicity, pharmacokinetic/pharmacology, previous human experience, and clinical development. The sponsor will then prepare a registration document that combines all the relevant data to allow the FDA to review and decide whether to approve marketing of the new drug. As discussed in Section 1.5, an NDA submission should include chemistry, pharmacology, toxicology, metabolism, manufacturing, quality controls, and clinical data along with the proposed labeling.

Good Clinical Practices

Good clinical practices (GCP) is usually referred to as a set of standards for clinical studies to achieve and maintain high-quality clinical research in a sensible and responsible manner.

The FDA, the Committee for Proprietary Medicinal Products (CPMP) for the European Community, the Ministry of Health and Welfare of Japan, and other countries worldwide

have each issued guidelines on good clinical practices. For example, the FDA promulgated a number of regulations and guidelines governing the conduct of clinical studies from which data will be used to support applications for marketing approval of drug products. The FDA regulations refer to those regulations specified in 21 CFR Parts 50, 56, 312, and 314, while the FDA guidelines are guidelines issued for different drug products such as *Guidelines for the Clinical Evaluation of Anti-Anginal Drugs* and *Guidelines for the Clinical Evaluation of Bronchodilator Drugs*. On the other hand, the European Community established the principles for their own GCP standard in all four phases of clinical investigation of medicinal products in July 1990. Basically these guidelines define the responsibilities of sponsors, monitors, and investigators in the initiation, conduct, documentation, and verification of clinical studies to establish the credibility of data and to protect the rights and integrity of study participants.

In essence GCP concerns patient protection and the quality of data used to prove the efficacy and safety of a drug product. GCP ensures that all data, information, and documents relating to a clinical study can be confirmed as being properly generated, recorded, and reported through the institution by independent audits. Therefore the basic GCP concerns are not only the protection of study subjects through informed consent and consultation by ethics committees such as IRB but also the responsibilities of the sponsors and monitors to establish written procedures for study monitoring and conduct and to ensure that such procedures are followed. In addition GCP emphasizes the responsibilities of investigator to conduct the study according to the protocol and joint responsibilities for data reporting, recording, analysis, and archiving as well as prompt reporting of serious adverse events. Moreover GCP calls for the most appropriate design for a valid statistical evaluation of the hypotheses of the clinical trials. The chosen design must suit the purpose with the best possible fit. Incorporating the concerns of GCP in the protocol will ensure a protocol of high standard, which in turn will help generate high-quality data.

Study conduct according to GCP standards requires regular visits to investigating center to monitor study progress. The activities of the sponsor's monitors that will affect the investigator and support staff should be stated in the protocol. Not only this is courteous, it prevents misunderstanding, facilities cooperation, and aids the speedy acquisition of completed case report form. The activities include frequency of monitoring visits, activities while on site (e.g., auditing CRFs), and departments to be visited (e.g., pharmacy). The practical effects of adopting GCP are that the investigator is audited by the sponsor's monitors (to confirm data on CRFs are a true transcript of original records), by a sponsor administratively separate from the clinical function and in some countries, by the national regulatory agency. The sponsor's monitors are audited by a compliance staff and by national regulatory agencies to confirm the accuracy of data recorded and the implementation of all written procedures such as standard operating procedure (SOP) and protocol.

Most of pharmaceutical companies and research institutions have a protocol review committee (PRC) to evaluate the quality and integrity of the protocol and hence to approve or disapprove the protocol. Some companies also ask the principal study medical monitor and statistician to submit a case report form (CRF) and a statistical analysis plan with mock tables and listing for presentation of the results to PRC at the same time when the protocol is submitted for review.

Lisook (1992) has assembled a GCP packet to assist the sponsors in the planning, execution, data analysis, and submission of results to the FDA. A summary of this GCP packet is given in Table 1.6.2. Most of these regulations have been discussed in the previous sections of this chapter. To improve the conduct and oversight of clinical research and to ensure the

Table 1.6.2 References to Keep at Hand for Good Clinical Practice

1. Information on FDA regulations
2. Center for Drug Evaluation and Research publications
3. Clinical Investigations (excerpt from the *Federal Register*, 9-27-1977)
4. Protection of Human Subjects, Informed Consent Forms
5. New Drug, Antibiotic, and Biologic Drug Product Regulations; Final Rule (excerpt from the *Federal Register*, 3-19-1987)
6. Investigational New Drug, Antibiotic, and Biologic Drug Product Regulations; Treatment Use and Sale; Final Rule (excerpt from the *Federal Register*, 5-22-1987)
7. *Guideline for the Monitoring of Clinical Investigations*
8. Investigational New Drug, Antibiotic, and Biologic Drug Product Regulations; Procedure Intended to Treat Life-Threatening and Severely Debilitating Illness; Interim Rule (excerpt from the *Federal Register*, 10-22-1988)
9. FDA IRB (Institution Review Board) Information Sheets
10. FDA Clinical Investigator Sheet
11. Reprint of Alan B. Lisook, M.D. FDA audits of clinical studies: Policy and procedure, *Journal of Clinical Pharmacology*, **30** (April 1990) 296-302.
12. Federal Policy for the Protection of Human Subjects; Notices and Rules (excerpt from the *Federal Register*, 6-18-1991)
13. *FDA Compliance Program Guidance Manual-Clinical Investigators* (10-1-1997)
14. *FDA Compliance Program Guidance Manual-Sponsors, Contract Research Organization and Monitors* (2-21-2001)
15. *FDA Compliance Program Guidance Manuals-Institutional Review Board* (10-1-1994)

protection of subjects participating in the FDA-regulated clinical research, the U.S. FDA established the Office of Good Clinical Practice (OGCP) within the Office of the Commissioned and its Office of Science Coordination and Communication in 2001. This new office has distinct roles from the Office of Human Research Protections (OHRP) of the Department of Health and Human Services (DHHS). These distinct roles include (1) coordination of the FDA's policies, (2) provision of leadership and direction through the administration of the FDA's Human Subject Protection/Good Clinical Practice Steering Committee, (3) coordination of the FDA's Bioresearch Monitoring program, (4) contribution to the international Good Clinical Practice harmonization activities, (5) planning and conducting training and outreach programs, and (6) serving as a liaison with OHRP and other federal agencies and other stakeholders committed to the protection of human research participants.

In the past, as demonstrated in Tables 1.5.2 and 1.5.3, Tables 1.5.5 and 1.5.6, health regulatory authorities in different countries have different requirements for approval of commercial use of the drug products. As a result, considerable resource had been spent by the pharmaceutical industry in the preparation of different documents for applications of the same pharmaceutical product to meet different regulatory requirements requested by different countries or regions. However, because of globalization of the pharmaceutical industry, arbitrary differences in regulations, increase of health care costs, need for reduction of time for patients to access new drugs, and of experimental use of humans and animals without compromising safety, the necessity to standardize these similar yet different regulatory requirements has been recognized by both regulatory authorities and pharmaceutical industry. Hence, The International Conference on Harmonization (ICH) of Technical Requirements for the Registration of Pharmaceuticals for Human Use was organized to provide an opportunity for important initiatives to be developed by regulatory authorities as

well as industry association for the promotion of international harmonization of regulatory requirements.

Currently, ICH, however, is only concerned with tripartite harmonization of technical requirements for the registration of pharmaceutical products among three regions: The European Union, Japan, and the United States. Basically, the organization of the ICH consists of two representatives, one from a regulatory authority and one from the pharmaceutical industry, from each of the three regions. As a result, the organization of the ICH consists of six parties of these three regions which include the European Commission of the European Union, the European Federation of Pharmaceutical Industries' Associations (EFPIA), the Japanese Ministry of Health, Labor and Welfare (MHLW), the Japanese Pharmaceutical Manufacturers Association (JPMA), the Centers for Drug Evaluation and Research and Biologics Evaluation and Research of the US FDA, and the Pharmaceutical Research and Manufacturers of America (PhRMA). The ICH steering committee was established in April, 1990 to (1) determine policies and procedures, (2) to select topics, (3) to monitor progress, and (4) to oversee preparation of biannual conferences. Each of the six parties has two seats on the ICH steering committee. The ICH steering committee also includes observers from the World Health Organization, the Canadian Health Protection Branch, and the European Free Trade Area which have one seat each on the committee. In addition, two seats of the ICH Steering Committee are given to the International Federation of Pharmaceutical Manufacturers Association (IFPMA), which represents the research-based pharmaceutical industry from 56 countries outside ICH regions. IFPMA also runs the ICH Secretariat at Geneva, Switzerland which coordinates the preparation of documentation.

In order to harmonize technical procedures the ICH has issued a number of guidelines and draft guidelines. After the ICH steering committee selected the topics, the ICH guidelines initiated by a concept paper and went through a 5-step review process given in Table 1.6.3. The number of ICH guidelines and draft guidelines at various stages of review process is given in Table 1.6.4. Table 1.6.5 provides a list of currently available ICH guidelines or draft guidelines pertaining to clinical trials while Table 1.6.6 gives the table of contents for the ICH draft guideline on general considerations for clinical trials. In addition, the table of contents of the ICH guidelines for good clinical practices: consolidated guidelines, for structure and content of clinical study reports, and for statistical principles for clinical trials are given, respectively in Tables 1.6.7, 1.6.8, and 1.6.9. From these tables, it can be seen that these guidelines are not only for harmonization of design, conduct, analysis, and report for a single clinical trial but also for consensus in protecting and maintaining the scientific integrity of the entire clinical development plan of a pharmaceutical entity. Along this line, Chow (1997, 2003) introduced the concept of *good statistics practice* (GSP) in drug development and regulatory approval process as the foundation of ICH GCP. The concepts and principles stated in the ICH clinical guidelines will be introduced, addressed, and discussed in the subsequent chapters of this book.

1.7 AIMS AND STRUCTURE OF THE BOOK

As indicated earlier, clinical trials are scientific investigations that examine and evaluate drug therapies in human subjects. Biostatistics has been recognized and extensively employed as an indispensable tool for planning, conduct, and interpretation of clinical trials. In clinical research and development the biostatistician plays an important role that

Table 1.6.3 Review Steps for the ICH Guidelines

<i>Step 1</i>
1. Harmonized topic identified
2. Expert working group (EWG) formed
3. Each party has a topic leader and a deputy
4. Rapporteur for EWR selected
5. Other parties represented on EWG as appropriate
6. Produce a guideline, policy statement, “points to consider”
7. Agreement on scientific issues
8. Sign-off and submit to the ICH steering committee
<i>Step 2</i>
1. Review of ICH document by steering committee
2. Sign-off by all six parties
3. Formal consultation in accord with regional requirements
<i>Step 3</i>
1. Regulatory rapporteur appointed
2. Collection and review of comments across all three regions
3. Step 2 draft revised
4. Sign-off by EWR regulatory members
<i>Step 4</i>
1. Forward to steering committee
2. Review and sign-off by three regulatory members of ICH
3. Recommend for adoption to regulatory bodies
<i>Step 5</i>
1. Recommendations are adopted by regulatory agencies
2. Incorporation into domestic regulations and guidelines

contributes toward the success of clinical trials. Well-prepared and open communication among clinicians, biostatisticians, and other related clinical research scientists will result in a successful clinical trial. Communication, however, is a two-way street: Not only (1) must the biostatistician effectively deliver statistical concepts and methodologies to his or her clinical colleagues but also (2) the clinicians must communicate thoroughly clinical and scientific principles embedded in clinical research to the biostatisticians. The biostatisticians

Table 1.6.4 Summary of the Number of ICH Guidelines or Draft Guidelines

	Step 1	Step 2	Step 3	Step 4	Step 5
Efficacy	0	0	0	0	15
Safety	0	0	1	0	13
Quality	0	0	6	0	17
Multidiscipline	0	0	1	0	4

Table 1.6.5 The ICH Clinical Guidelines or Draft Guidelines

-
1. E1A: The Extent of Population Exposure to Assess Clinical Safety for Drugs Intended for Long-term treatment of Non-Life-Threatening Conditions
 2. E2A: Clinical Safety Data Management: Definitions and Standards for Expedited Reporting
 3. E2B: Data Elements for Transmission of Individual Case Safety Reports
 4. E2B(M): Data Elements for Transmission of Individual Case Safety Reports
 5. M2:E2B(M): Electronic Transmission of Individual Case Safety Reports Message Specification
 6. E2C: Clinical Safety Data Management: Periodic Safety Update Reports for Marketed Drugs
 7. E3: Structure and Content of Clinical Studies
 8. E4: Dose-Response Information to Support Drug Registration
 9. E5: Ethnic Factors in the Acceptability of Foreign Clinical Data
 10. E6: Good Clinical Practice: Consolidated Guideline
 11. E7: Studies in Support of Special Populations: Geriatrics
 12. E8: General Considerations for Clinical Trials
 13. E9: Statistical Principles for Clinical Trials
 14. E10: Choice of Control Group in Clinical Trials
 15. E11: Clinical Investigation of Medicinal Products in the Pediatric Population
 16. M4: Common Technical Document for the registration of Pharmaceuticals for Human Use
 - Main Document (Organization)
 - Efficacy
 - Safety
 - Safety Appendices
 - Quality
 17. Principles for Clinical Evaluation of New Antihypotensive Drugs
 18. Draft Guidelines M2: Electronic Technical Document Specification (eTD)
-

Table 1.6.6 The Table of Contents for the Guideline on General Considerations for Clinical Trials

-
1. Objectives of this document
 2. General principles
 - 2.1 Protection of clinical trial subjects
 - 2.2 Scientific approach in design and analysis
 3. Development methodology
 - 3.1 Considerations for development
 - 3.1.1 Nonclinical studies
 - 3.1.2 Quality of investigational medicinal products
 - 3.1.3 Phases of clinical development
 - 3.1.4 Special considerations
 - 3.2 Considerations for individual clinical trials
 - 3.2.1 Objectives
 - 3.2.2 Design
 - 3.2.3 Conduct
 - 3.2.4 Analysis
 - 3.2.5 Reporting
-

**Table 1.6.7 Table of Contents for the ICH Guideline on Good Clinical Practice:
Consolidated Guideline**

- Introduction
- 1. Glossary
- 2. The Principles of ICH GCP
- 3. The Institutional Review Board/Independent Ethnic Committee (IRB/IEC)
 - 3.1 Responsibilities
 - 3.2 Composition, functions, and operations
 - 3.3 Procedures
 - 3.4 Records
- 4. Investigators
 - 4.1 Investigator's qualifications and agreements
 - 4.2 Adequate resources
 - 4.3 Medical care of trial subjects
 - 4.4 Communication with IRB/IEC
 - 4.5 Compliance with protocol
 - 4.6 Investigational products
 - 4.7 Randomization procedures and unblinding
 - 4.8 Informed consent of trial subjects
 - 4.9 Records and reports
 - 4.10 Progress reports
 - 4.11 Safety reporting
 - 4.12 Premature termination or suspension of a trial
 - 4.13 Final report(s) by investigator/institution
- 5. Sponsor
 - 5.1 Quality assurance and quality control
 - 5.2 Contract research organization
 - 5.3 Medical expertise
 - 5.4 Trial design
 - 5.5 Trial management, data handling, recordingkeeping, and independent data monitoring committee
 - 5.6 Investigator selection
 - 5.7 Allocation of duties and functions
 - 5.8 Compensation to subjects and investigators
 - 5.9 Financing
 - 5.10 Notification/submission to regulatory authority(ies)
 - 5.11 Confirmation of review of IRE/IEC
 - 5.12 Information on investigational product(s)
 - 5.13 Manufacturing, packaging, labeling, coding investigation product(s)
 - 5.14 Supplying and handling, investigational product(s)
 - 5.15 Record access
 - 5.16 Safety information
 - 5.17 Adverse drug reaction reporting
 - 5.18 Monitoring
 - 5.19 Audit
 - 5.20 Noncompliance
 - 5.21 Premature termination or suspension of a trial
 - 5.22 Clinical trial/study reports
 - 5.23 Multicenter trials
- 6. Clinical Trial Protocol and Protocol Amendment(s)
 - 6.1 General information
 - 6.2 Background information

Table 1.6.7 (Continued)

-
6. Clinical Trial Protocol and Protocol Amendment(s) (*Continued*)
- 6.3 Trial objectives and purpose
 - 6.4 Trial design
 - 6.5 Selection and withdrawal of subjects
 - 6.6 Treatment of subjects
 - 6.7 Assessment of efficacy
 - 6.8 Assessment of safety
 - 6.9 Statistics
 - 6.10 Direct assess to source data/documents
 - 6.11 Quality control and quality assurance
 - 6.12 Ethics
 - 6.13 Data handling and recordkeeping
 - 6.14 Financing and insurance
 - 6.15 Publication
 - 6.16 Supplements
7. Investigator's Brochure
- 7.1 Introduction
 - 7.2 General considerations
 - 7.3 Contents of the investigator's brochure
 - 7.4 Appendix 1
 - 7.5 Appendix 2
8. Essential documents for the conduct of a clinical trial
- 8.1 Introduction
 - 8.2 Before the clinical phase of the trial commences
 - 8.3 During the clinical conduct of the trial
 - 8.4 After completion or termination of the trial
-

Table 1.6.8 Table of Contents for the ICH Guideline on Structure and Contents of Clinical Study Reports

-
- Introduction to the guideline
1. Title page
 2. Synopsis
 3. Table of contents for the individual clinical study report
 4. List of abbreviations and definition of terms
 5. Ethics
 6. Investigators and study administrative structure
 7. Introduction
 8. Study objectives
 9. Investigational plan
 10. Study patients
 11. Efficacy evaluation
 12. Safety evaluation
 13. Discussion and overall conclusions
 14. Tables, figures, graphs referred to but not included in the text
 15. Reference list
 16. Appendices
-

Table 1.6.9 Table of Contents for the ICH Guideline on Statistical Principles for Clinical Trials

I.	Introduction
1.1	Background and purpose
1.2	Scope and purpose
II.	Considerations for Overall Clinical Development
2.1	Trial content
2.2	Scope of trials
2.3	Trial techniques to avoid bias
III.	Trial Design Considerations
3.1	Design configuration
3.2	Multicenter trials
3.3	Type of comparison
3.4	Group sequential designs
3.5	Sample size
3.6	Data capture and processing
IV.	Trial Conduct Considerations
4.1	Trial monitoring and interim analysis
4.2	Changes in inclusion and exclusion criteria
4.3	Accrual rates
4.4	Sample size adjustment
4.5	Interim analysis and early stopping
4.6	Role of independent data monitoring committee (IDMC)
V.	Data Analysis Considerations
5.1	Prespecification of the analysis
5.2	Analysis sets
5.3	Missing values and outliers
5.4	Data transformation
5.5	Estimation, confidence interval, and hypothesis testing
5.6	Adjustment of significance and confidence levels
5.7	Subgroups, interaction, and covariates
5.8	Integrity of data and computer software validity
VI.	Evaluation of Safety and Tolerability
6.1	Scope of evaluation
6.2	Choice of variables and data collection
6.3	Set of subjects to be evaluated and presentation of data
6.4	Statistical evaluation
6.5	Integrated summary
VII.	Reporting
7.1	Evaluation and reporting
7.2	Summarizing the clinical database

Annex I Glossary

can then formulate these clinical and scientific principles into valid statistical hypotheses under an appropriate statistical model. Overall, the integrity, quality, and success of a clinical trial depends on the interaction, mutual respect, and understanding between the clinicians and the biostatisticians.

The aim of this book is not only to fill the gap between clinical and statistical disciplines but also to provide a comprehensive and unified presentation of clinical and scientific issues, statistical concepts, and methodology. Moreover the book will focus on the interactions

between clinicians and biostatisticians that often occur during various phases of clinical research and development. This book is also intended to give a well-balanced summarization of current and emerging clinical issues and recently developed corresponding statistical methodologies. Although this book is written from the viewpoint of pharmaceutical research and development, the principles and concepts presented in this book can also be applied to a nonbiopharmaceutical setting.

It is our goal to provide a comprehensive reference book for physicians, clinical researchers, pharmaceutical scientists, clinical or medical research associates, clinical programmers or data coordinators, and biostatisticians or statisticians in the areas of clinical research and development, regulatory agencies, and academe.

The scope of this book covers clinical issues, which may occur during various phases of clinical trials in pharmaceutical research and development, their corresponding statistical interpretations, concepts, designs and analyses, which are adopted to address these important clinical issues. Basically, this book is devoted to the concepts and methodologies of design and analysis of clinical trials. As a result, this book can be divided into two parts: concepts and methodologies. Each part consists of several chapters with different topics. Each part and each chapter are self-contained. But, at the same time, parts and chapters are arranged in a sensible manner such that there is a smooth transition between parts and from chapter to chapter within each part.

Chapter 1 provides an overview of clinical development for pharmaceutical entities, drug research and development process in the pharmaceutical industry, regulatory review, and approval processes and requirements. Also included in this chapter are the aim and structure of the book. Chapters 2 to 7 cover the concepts of design and analysis of clinical trials. Chapter 2 introduces basic statistical concepts such as uncertainty, bias, variability, confounding, interaction, clinical significance and equivalence, and reproducibility and generalizability. Chapter 3 provides some fundamental considerations for choosing a valid and suitable design for achieving study objectives of clinical trials under various circumstances. Chapter 4 illustrates the concepts and different methods of randomization and blinding, which are critically indispensable for the success and integrity of clinical trials. Chapter 5 introduces different types of statistical designs for clinical trials. These study designs include parallel group, crossover, titration, enrichment, clustered, group-sequential, placebo-challenging, and blinder-reader designs. Also included in this chapter is the discussion of the relative merits and disadvantages of these study designs. Specific designs for cancer clinical trials are introduced in Chapter 6. These designs include standard escalation, accelerated titration, and continual reassessment method (CRM) in determination of maximum tolerable dose (MTD) for phase I cancer trials. In addition, Simon's optimal two-stage design and randomized phase II designs are also discussed. Various types of clinical trials, including multicenter, superiority, dose-response, active control, equivalence and noninferiority, drug-to-drug interaction, combination, and bridging trials, are discussed in Chapter 7.

Chapters 8 through 13 cover methodologies and various issues that are commonly encountered in the analysis of clinical data. As clinical endpoints can generally be classified into three types: continuous, categorical, and censored data, different statistical methods for analysis of these three types of clinical data are necessary. Chapters 8, 9, and 10 discuss the advantages and limitations of statistical methods for analysis of continuous, categorical, and censored data, respectively. In addition, group sequential procedures for interim analysis are also given in Chapter 10. Chapter 11 provides different procedures for sample size calculation for various types of data under different study designs. Chapter 12

discusses statistical issues in analyzing efficacy data. These issues include baseline comparison, intention-to-treat analysis versus evaluable or per-protocol analysis, adjustment of covariates, multiplicity, the use of genomic information for assessment of efficacy, and data monitoring. Chapter 13 focuses on the issues for analysis of safety data, which include the extent of exposure, coding and analysis of adverse events, the analysis of laboratory data, and the use of genomic information for evaluation of drug safety.

Issues of study protocols and clinical data management are provided in Chapters 14 and 15, respectively. Chapter 14 focuses on the development of a clinical protocol. This chapter discusses the structure and components of an adequate and well-controlled clinical trial protocol, issues that are commonly encountered in protocol development, commonly seen deviations in the conduct of a clinical trial, clinical monitoring, regulatory audit and inspection, and assessment of the quality and integrity of clinical trials. Chapter 15 summarizes basic standard operating procedures for good clinical data management practice. These standard operating procedures cover the development of case report forms (CRF), database development and validation, data entry, validation and correction, database finalization and lock, CRF flow and tracking, and the assessment of clinical data quality.

For each chapter, whenever possible, real examples from clinical trials are included to demonstrate the clinical and statistical concepts, interpretations, and their relationships and interactions. Comparisons regarding the relative merits and disadvantages of the statistical methodology for addressing different clinical issues in various therapeutic areas are discussed wherever deemed appropriate. In addition, if applicable, topics for future research development are provided.

2

BASIC STATISTICAL CONCEPTS

2.1 INTRODUCTION

As was indicated in the preceding chapter, in general, the FDA requires that two adequate well-controlled clinical trials be conducted to demonstrate the effectiveness and safety of a drug product. The success of an adequate well-controlled clinical trial depends on a well-designed protocol. A well-designed protocol describes how the clinical trial is to be carried out, which ensures the quality of clinical data collected from the trial. Based on the high-quality clinical data, appropriate statistical methods can then be applied to provide a valid and unbiased assessment of the efficacy and safety of the drug product. Spilker (1991) indicated that the greater the attention paid to the planning phase of a clinical trial, the greater the likelihood that the clinical trial will be conducted as desired. In this chapter, we will describe several basic statistical concepts and issues that have a great impact on the success of a clinical trial during its planning, design, execution, analysis, and reporting phases. These basic statistical concepts include uncertainty and probability, bias and variability, confounding and interaction, and descriptive and inferential statistics using hypotheses testing and p -values, for example.

In the medical community, there are many unknowns remaining in the clinical research of certain diseases such as AIDS. These unknowns or uncertainties are often scientific questions of particular interest to clinical scientists. Once a scientific question regarding the uncertainty of interest is clearly stated, clinical trials are necessarily conducted to provide scientific or clinical evidence to statistically address the uncertainty. Under some underlying probability distribution assumption, a statistical inference can then be derived based on clinical data collected from a representative sample of the targeted patient population. To provide a valid statistical assessment of the uncertainty with a desired accuracy and reliability, statistical and/or estimation procedures should possess the properties of

Design and Analysis of Clinical Trials: Concepts and Methodologies, Second Edition

By Shein-Chung Chow and Jen-pei Liu

ISBN 0-471-24985-8 Copyright © 2004 John Wiley & Sons, Inc.

unbiasedness and least variability whenever possible. In practice, well-planned statistical designs can generally serve to avoid unnecessary bias and minimize the potential variability that can occur during the conduct of the clinical trials. In some clinical trials, design factors such as race and gender may have an impact on the statistical inference of clinical evaluation of the study medication. In this case, it is suggested that possible confounding and/or interaction effects be carefully identified and separated from the treatment effect in order to have a valid and unbiased assessment of the clinical evaluation of the study medication. After the completion of a clinical trial, the collected clinical data can either be summarized descriptively to provide a quick overview of clinical results or be analyzed to provide statistical inference on clinical endpoints of interest. Descriptive statistics usually provide useful information regarding a potential treatment effect. This information can be confirmed by a valid statistical inference with a certain assurance that can be obtained through an appropriate statistical analysis such as hypotheses testing and *p*-values.

These basic statistical concepts play an important role in the success of clinical trials. They are helpful not only at the very early stage of a study's concept statement development but also at the stage of protocol development. In addition to these basic statistical concepts, there are many statistical/medical issues that can affect the success of a clinical trial. For example, clinical scientists may be interested in establishing clinical efficacy using a one-sided test procedure rather than a two-sided test procedure based on prior experience of the study medication. It should be noted that different test procedures require different sample sizes and address different kinds of uncertainty regarding the study medication. In addition, clinical scientists always focus on clinical difference rather than statistical difference. It should be noted that the statistical test is meant to detect a statistical difference with a desired power. The discrepancy between a clinical difference and a statistical difference has an impact on the establishment of clinical equivalence between treatments. As a result, these issues are also critical in the success of clinical trials which often involve considerations from different perspectives such as political, medical, marketing, regulatory, and statistical.

2.2 UNCERTAINTY AND PROBABILITY

For a medication under investigation, there are usually many questions regarding the properties of the medication that are of particular interest to clinical scientists. For example, the clinical scientists are interested to know whether the study medication works for the intended indication and patient population. In addition, the clinical scientists may be interested in knowing whether the study medication can be used as a substitute for other medications currently available on the market. To address these questions (or uncertainties) regarding the study medication, clinical trials are necessarily conducted to provide scientific/clinical evidence for a fair scientific/clinical evaluation/justification. In order to address the uncertainty regarding the study medication for the targeted patient population, a representative sample is typically drawn for clinical evaluation according to a well-designed protocol. Based on clinical results from the study, statistical inference on the uncertainty can then be made under some underlying probability distribution assumption. As a result, the concept of uncertainty and probability plays an important role for clinical evaluation of a study medication.

Uncertainty

Bailar (1992) indicated that uncertainties of interest to clinical scientists include uncertainty from confounders, uncertainty regarding scientific or medical assumptions (or

hypotheses), and uncertainty about the generalization of the results from animals to humans. These uncertainties are to be verified through clinical trials. A recent example of uncertainty on medical assumptions is the cardiac arrhythmia suppression trial (CAST, 1989). Two antiarrhythmic agents, namely, encainide and flecainide were approved for the indication of ventricular arrhythmia by the FDA based on indisputable evidence on the suppression of objective endpoint premature ventricular beats (PVB) per hours as documented by ambulatory 24-hour Holter monitor. This is because the occurrence of premature ventricular depolarization is considered a risk factor in the survivor of myocardial infarction. The approval of encainide and flecainide in treating patients who survived myocardial infarction is based on the fundamental assumption that the suppression of PVB will reduce the chance of subsequent sudden death. This crucial assumption of treating patients with asymptomatic or symptomatic ventricular arrhythmia after myocardial infarction was never challenged until the CAST trial initiated by the U.S. National Heart, Lung, and Blood Institute. After an average of 10 months of follow-up, an interim analysis conducted by the investigators of the CAST discovered that the relative risk of deaths from arrhythmia and nonfatal cardiac arrest of the patients receiving encainide or flecainide (33 of 725) as compared to the placebo (9 of 725) is 3.5. In other words, the chance of death or suffering cardiac arrest for patients who took either encainide or flecainide were three times as high as those who took placebo. The study was terminated shortly after this finding. As a result, the investigators recommended that neither encainide nor flecainide be used to treat patients with asymptomatic or mildly symptomatic ventricular arrhythmia after myocardial infarction. The saga of CAST demonstrates that the assumption of the suppression of PVB as a surrogate clinical endpoint for survival of this patient population is inadequate. In clinical trials, we usually make scientific or medical assumptions that are based on previous animal or human experiences. It is suggested that these critical assumptions be precisely stated in the protocol for clinical test, evaluation, and interpretation.

Uncertainty regarding confounders and scientific or medical assumptions can not usually be quantified until the confounders are controlled or assumptions are properly investigated. Note that the concept of confounders will be introduced later in this chapter. In practice, the uncertainty caused by known variations can be statistically quantified. For example, in the GUSTO I (the Global Use of Strategies to Open Occluded Coronary Arteries, 1993) study it was suggested that for the patients with evolving myocardial infarction, on the average, the intravenous administration of t-PA (accelerated) over a period of one and a half hour produced a 14.5% reduction in 30-day mortality as compared with the streptokinase therapy. An interesting question is, How likely is it that the same mortality reduction will be observed in patients with similar characteristics as the 41,021 patients enrolled in the GUSTO I study? It depends on underlying source of variations. The accelerated t-PA provided a 21% mortality reduction for patients who are younger than 75 years old, while only a 9% reduction in the patients older than 75. This illustrates the uncertainty due to biological variation. Thus, due to the various sources of variation, before the administration of the accelerated t-PA, cardiologists can only expect that on the average, a 93.7% myocardial infarction will be saved. The 30-day survival of a patient can be realized by the accelerated t-PA only after until it is administrated and the patient is observed over a period of 30 days. Even though the relationship between the accelerated t-PA and 30-day mortality was deterministic due to the variation from different causes, clinicians have to think probabilistically in the application of any clinical data, such as the results of the GUSTO I study.

Probability

The purpose of clinical trials is not only to investigate or verify some scientific or medical hypotheses of certain interventions in a group of patients but also to be able to apply the results to the targeted patient population with similar characteristics. This process is called (statistical) inference. It is the process by which clinicians can draw conclusions based on the results observed from the targeted patient population. Suppose that a clinical trial is planned to study the effectiveness of a newly developed cholesterol-lowering agent in patients with hypercholesterolemia as defined by nonfasting plasma total cholesterol level being greater than 250 mg/dL. Furthermore, assume that this new agent is extremely promising as shown by previous small studies and that the elevation of the cholesterol level is a critical factor for reduction of the incidence of coronary heart disease. For this reason, the government is willing to provide unlimited resources so that every patient with hypercholesterolemia in the country has the opportunity to be treated with this promising agent. One of the primary clinical responses is the mean reduction in total cholesterol level after six months of treatment from the baseline. In this hypothetical trial, the targeted patient population is patients with hypercholesterolemia. The mean reduction of total cholesterol level after six months of treatment is a characteristic regarding the targeted patient population that we are interested in this study. If we measure mean reduction in total cholesterol level for each patient, the mean reductions in total cholesterol level from baseline would form a population distribution. Statistically, a distribution can be characterized by its location, spread and skewness. The location of a distribution is also referred to as the central tendency of the distribution. The most commonly used measures for the central tendency are arithmetic mean, median, and mode. The *arithmetic mean* is defined as the sum of the reductions divided by the number of patients in the patient population. The most frequently occurring reductions are called the *mode*, while the median is the middle value of the reductions among all patients. In other words, half of the patients have their reductions above the median and the other half have theirs below it. *Spread* is the variation or dispersion among the patients. The commonly used measures for variation are range, variance, and standard deviation (SD). The *range* is simply the difference between the largest and smallest reductions in the patient population. The *variance*, however, is the sum of the squares of the deviations from the mean divided by the number of the patients. The most commonly employed measure for dispersion is the *standard deviation*, which is defined as the positive square root of the variance. These measures for description of the population distribution are called parameters. Statistically, we can impose some probability laws to describe the population distribution. The most important and frequently used probability law is probably the bell-shaped normal distribution which serves an adequate and satisfactory model for description of many responses in clinical research such as height, weight, total cholesterol level, blood pressure, and many others. For this hypothetical study, suppose that mean and standard deviation of the reduction in total cholesterol level from baseline are 50 and 10 mg/dL, respectively. Then about 68% patients are expected to have reductions between 40 and 60 mg/dL. The chance that a reduction exceeds 70 mg/dL is only about 2.5%. Under the normal probability law, the uncertainty of reduction in total cholesterol level can be completely quantified because the scope of this trial is the entire population of patients with hypercholesterolemia. In reality, however, the government will have a limited budget/resource. We therefore could not conduct such intensive study. Alternatively, we randomly select a representative sample from the patient population. For this sample, similar descriptive measures for central tendency and variation are computed.

They are called *statistics*. These statistics are estimates for the corresponding parameters of the patient population. Since it is almost impossible to conduct a clinical trial on the entire patient population and the parameters are always unknown, statistics, in turn, can be used to approximate its corresponding population parameters. This process of making a definitive conclusion about the patient population based on the results of randomly selected samples is referred to as *statistical inference*. Statistical inference provides an approximation to the parameters of the patient population with certain assurance. It therefore involves uncertainty too. The closeness of the approximation to the unknown population parameters by the sample statistics can also be quantified by the probability laws in statistics. These probability laws are called the sampling distribution of sample statistics. The sampling distributions are the basis of statistical inference. To provide a better understanding, we continue the above example concerning the study of reduction in total cholesterol level in patients with hypercholesterolemia. Suppose we randomly select a sample of 100 patients from the patient population and calculate the mean reduction after six months of treatment. In practice, we can select another sample of 100 patients from the same patient population with replacement and compute its mean reduction. We can repeat this sampling process indefinitely and compute mean reduction for each sample drawn from the same population. The sampling distribution for the sample mean reductions in total cholesterol level for the sample size of 100 patients can then be determined as the frequency distribution of these mean reductions. It can be verified that the mean of the sampling distribution of the sample mean reductions in total cholesterol level is equal to its unknown population parameter. This desirable statistical property is called *unbiasedness*. In practice, we only draw one sample from the patient population, and hence we only have one sample mean cholesterol reduction. It is then of interest to know how we would judge the closeness of the sample mean cholesterol reduction to its corresponding population parameter. This can be determined by the precision of the sample mean cholesterol reduction. With the sample size of 100, therefore, the variance of the sample mean is simply equal to the population variance divided by 100. The square root of the variance of the sample mean is called the *standard error* (SE) of the sample mean. The approximation of the population mean by the sample mean can be quantified by its standard error. The smaller the standard error is, the closer to the unknown population mean it is. When the sample size is sufficiently large, the sampling distribution will behave like a normal distribution regardless of its corresponding population distribution. This important property is called the *Central Limit Theorem*. As a result we can quantify the standard error of the sample mean in conjunction with the central limit theorem in terms of probability, that is, the closeness of the approximation of the sample mean to the population mean in the total cholesterol reduction provided that the sample size is at least of moderate size.

2.3 BIAS AND VARIABILITY

As was indicated earlier, the FDA requires that the results from clinical trials be accurate and reliable in order to provide a valid and unbiased assessment of true efficacy and safety of the study medication. The accuracy and reliability are usually referred to as the closeness and the degree of the closeness of the clinical results to the true value regarding the targeted patient population. The accuracy and reliability can be assessed by the bias and variability of the primary clinical endpoint used for clinical assessment of the study medication. In what follows, we will provide more insight regarding bias and variability separately.

Bias

Since the accuracy of the clinical results is referred to as closeness to the true value, we measure any deviation from the true value. The deviation from the true value is considered as a *bias*. In clinical trials, clinical scientists would make any attempt to avoid bias in order to ensure that the collected clinical results are accurate. It, however, should be noted that most biases are probably caused by human errors. *Webster's II New Riverside University Dictionary* (1984) defines bias as an inclination or preference, namely one that interferes with impartial judgment. Along this line, Minert (1986) considers bias as a preconceived personal preference or inclination that influences the way in which a measurement, analysis, assessment, or procedure is performed or reported. Spilker (1991), on the other hand, views bias as a systematic error that enters a clinical trial and distorts the data obtained, as opposed to a random error that might enter a clinical trial. Yet another definition, by Sackett (1979), describes bias as any process at any stage of inference that tends to produce results or conclusions that differ systematically from the truth. ICH E9 guideline, entitled *Statistical Principles for Clinical Trials*, defines bias as the systematic tendency of any factors associated with the design, conduct, analysis, and evaluation of the results of clinical trials to make the estimate of a treatment effect deviate from its true value. Thus, the bias could occur at any stage of a clinical trial, and it mainly comes from the four sources of design, conduct, analysis, and evaluation of results. Bias caused by the deviation in conduct is referred to as *operational bias*. Bias introduced from the other sources is referred to as *statistical bias*. As a summary of these different definitions, we define bias as a systematic error that deviates data from the truth caused by the partial judgment or personal preference that can occur at any stage of a clinical trial.

Clinical trials are usually planned, designed, executed, analyzed, and reported by a team that consists of clinical scientists from different disciplines to evaluate the effects of the treatments in a targeted population of human subjects. When there are such nonnegligible differences in human background, education, training, and opinions, it is extremely difficult to remain totally impartial to every aspect at all stages of a clinical trial. Bias inevitably occurs. Where bias occurs, the true effects of the treatment cannot be accurately estimated from the collected data. Since it is almost impossible for a clinical trial to be free of any biases, it is crucial to identify any potential bias that may occur at every stage of a clinical trial. Once the potential bias are identified, one can then implement some procedures such as blinding or randomization to minimize or eliminate the bias.

Sackett (1979) partitions a clinical trial (or research) into seven stages at which bias can occur. These seven stages are (1) in reading up on the field, (2) in specifying and selecting the study sample, (3) in executing the experimental maneuver, (4) in measuring exposures and outcomes, (5) in analyzing the data, (6) in interpreting the analysis result, and (7) in publishing the results. Sackett also provides a detailed catalog of biases for each stage of a case-control trial (see Table 2.3.1). Although the purpose of this catalog is to demonstrate that more biases can occur in case-control studies than in randomized control trials, it still can be applied to the different phases of clinical trials. In addition to the 57 types of bias listed by Sackett (1979), Spilker (1991) describes the six types of biases summarized in Table 2.3.2. Here we will classify all the possible biases into three groups: bias due to selection, observation, and statistical procedures.

Selection bias is probably the most common source of bias that can occur in clinical trials. For example, at the planning stage of a clinical trial, selection bias can occur if clinical scientists review only a partial existing literature on current treatments for a certain

Table 2.3.1 Catalog of Biases*1. In Reading the Literature*

- 1.1 *Bias of rhetoric.* Any of several techniques used to convince the reader without appealing to reasons.
- 1.2 *All's well literature bias.* Scientific or professional societies may publish reports or editorials that omit or play down controversies or disparate results.
- 1.3 *One-sided reference bias.* Authors may restrict their references to only those works that support their position; a literature review with a single starting point risks confinement to a single side of the issue.
- 1.4 *Positive results bias.* Authors are more likely to submit, and editors accept, positive than null results.
- 1.5 *Hot stuff bias.* When a topic is hot, neither investigators or editors may be able to resist the temptation to publish additional results, no matter how preliminary or shaky.

2. In Specifying and Selecting the Study Sample

- 2.1 *Popularity bias.* The admission of patients to some practices, institutions, or procedures (surgery, autopsy) is influenced by the interest stirred by the presenting and its certain causes.
- 2.2 *Centripetal bias.* Reputations of certain clinicians and institutions cause individuals with certain disorders or exposures to gravitate toward them.
- 2.3 *Referral filter bias.* As a group of ill persons referred from primary to secondary to tertiary care, the number of rare causes, multiple diagnoses, and hopeless cases may increase.
- 2.4 *Diagnostic access bias.* Individuals differ in their geographic, temporal, and economic access to the diagnostic procedures which label them as having a given disease.
- 2.5 *Diagnostic suspicion bias.* Knowledge of the subject's prior exposure to a putative cause (ethnicity, taking a certain drug, having a second disorder, being exposed in an epidemic) may influence both the intensity and the outcome of the diagnostic procedure.
- 2.6 *Unmasking (detection signal) bias.* An innocent exposure may become suspect if, rather than causing a disease, it causes a sign or symptom that precipitates a search for the disease, such as the current controversy over postmenopausal estrogens and cancer of the endometrium.
- 2.7 *Mimicry bias.* An innocent exposure may become suspect if, rather than causing a disease, it causes a (benign) disorder that resembles the disease.
- 2.8 *Previous opinion bias.* The tactics and results of a previous diagnostic process on a patient, if known, may affect the tactics and results of a subsequent diagnostic process on the same patient, such as multiple referrals among hypertensive patients.
- 2.9 *Wrong sample size bias.* Samples that are too small can prove noting; samples that are too large can prove anything.
- 2.10 *Admission rate (Berkson) bias.* If hospitalization rates differ for different exposure/disease groups, the relation between exposure and disease will become distorted in hospital-based studies.
- 2.11 *Prevalence-incidence (Neyman) bias.* A late look at those exposed (or affected) early will miss fatal and other short episodes, plus mild or "silent" cases and cases where evidence of exposure disappears with disease onset.
- 2.12 *Diagnostic vogue bias.* The same illness may receive different diagnostic labels at different points in space or time.
- 2.13 *Diagnostic purity bias.* "Pure" diagnostic groups that exclude comorbidity may become nonrepresentative.
- 2.14 *Procedure selection bias.* Certain clinical procedures may be preferentially offered to those who are poor risk, such as selection of patients for "medical" versus "surgical" therapy.
- 2.15 *Missing clinical data bias.* Clinical data may be missing if they are normal, negative, never measured, or measured but never recorded.

Table 2.3.1 (Continued)

-
- 2.16 *Noncontemporaneous control bias.* Secular changes in definitions, exposures, diagnoses, diseases, and treatments may render noncontemporaneous controls noncomparable.
- 2.17 *Starting time bias.* Failure to identify a common starting time for exposure or illness may lead to systematic misclassification.
- 2.18 *Unacceptable disease bias.* Socially unacceptable disorders (V.D., suicide, insanity) tend to be under-reported.
- 2.19 *Migrant bias.* Migrants may differ systematically from those who stay home.
- 2.20 *Membership bias.* Membership in a group (the employed, joggers, etc.) may imply a degree of health that differs systematically from the general population, such as more exercise and recurrent myocardial infarction.
- 2.21 *Nonrespondent bias.* Nonrespondents (or “latecomers”) from a specified sample may exhibit exposures or outcomes that differ from those of early respondents such as in cigarette smoking.
- 2.22 *Volunteer bias.* Volunteers in a study sample may exhibit exposures or outcomes (they tend to be healthier) that differ from those of nonvolunteers or latecomers, such as in the screening selection.

3. In Executing the Experiment Maneuver (or Exposure)

- 3.1 *Contamination bias.* In an experiment when members of the control group inadvertently receive the experiment maneuver, the difference in outcomes between experimental and control patients may be systematically reduced.
- 3.2 *Withdrawal bias.* Patients who are withdrawn from an experiment may differ systematically from those who remain, such as in a neurosurgical trial of surgical versus medical therapy of cerebrovascular disease, patients who died during surgery were withdrawn as “unavailable for follow-up” and excluded from early analyses.
- 3.3 *Compliance bias.* In requiring patient adherence to therapy, issues of efficacy become confounded with those of compliance.
- 3.5 *Bogus control bias.* When patients allocated to an experimental maneuver die or sicken before or during its administration and are omitted or reallocated to the control group, the experimental maneuver will appear spuriously superior.

4. In Measuring Exposure and Outcomes

- 4.1 *Insensitive exposures and outcomes.* when outcome measures are incapable of detecting clinically significant changes or difference, type II errors occur.
- 4.2 *Underlying causing bias (rumination bias).* Patients may ruminate about possible causes for their illness and exhibit different recall or prior exposure than controls.
- 4.3 *End-digit preference bias.* In converting analog to digital data, observers may record some terminal digits with an unusual frequency.
- 4.4 *Apprehension bias.* Certain measures (pulse, blood pressure) may alter systematically from their usual levels when the subject is apprehensive.
- 4.5 *Unacceptability bias.* Measurements that hurt, embarrass, or invade privacy may be systematically refused or evaded.
- 4.6 *Obsequiousness bias.* Subjects may systematically alter questionnaire responses in the direction they perceive desired by the investigator.
- 4.7 *Expectation bias.* Observers may systematically err in measuring and recording observations so that they concur with prior expectations.
- 4.8 *Substitution game.* The substitution of a risk factor that has not been established as causal for its associated outcome.
- 4.9 *Family information bias.* The flow of family information about exposure and illness is stimulated by, and directed to, a newly arising case.

Table 2.3.1 (Continued)

-
- 4.10 *Exposure suspicion bias.* A knowledge of the subject's disease status may influence both the intensity and outcome of a search for exposure to the putative cause.
- 4.11 *Recall bias.* Questions about specific exposures may be asked several times of cases but only once of controls.
- 4.12 *Attention bias.* Study subjects may systematically alter their behavior when they know they are being observed.
- 4.13 *Instrument bias.* Defects in the calibration or maintenance of measurement instruments may lead to systematic deviations from true value.

5. In Analyzing Data

- 5.1 *Post hoc significant bias.* When decision levels or "tails" for α and β are selected after the data have been examined, conclusions may be biased.
- 5.2 *Data dredging bias (looking for the pony).* When data are reviewed for all possible associations without prior hypothesis, the results are suitable for hypothesis-forming activities only.
- 5.3 *Scale degradation bias.* The degradation and collapsing of measurement scales tend to obscure differences between groups under comparison.
- 5.4 *Tidying-up bias.* The exclusion of outliers or other untidy results cannot be justified on statistical grounds and may lead to bias.
- 5.5 *Reported peek bias.* Repeated peeks at accumulating data in a randomized trial are not independent and may lead to inappropriate termination.

6. In Interpreting the Analysis

- 6.1 *Mistaken identity bias.* In compliance trials, strategies directed toward improving the patients' compliance may, instead or in addition, cause the treating clinician to prescribe more vigorously such that the effect on achievement of the treatment goal may be misinterpreted.
- 6.2 *Cognitive dissonance bias.* The belief in a given mechanism may increase rather than decrease in the face of contradictory evidence.
- 6.3 *Magnitude bias.* In interpreting a finding, the selection of a scale of measurement may markedly affect the interpretation; for example, 1,000,000 may be also be 0.0003% of the national budget.
- 6.4 *Significance bias.* The confusion of statistical significance with biological or clinical or health care significance can lead to fruitless and useless conclusion.
- 6.5 *Correlation bias.* Equating correlation with causation leads to errors of both kinds.
- 6.6 *Underexhaustion bias.* The failure to exhaust the hypothesis space may lead to authoritarian rather than authoritative interpretation.
-

Source: Sackett (1979).

disease. A review that does not provide a full spectrum of all possible positive and negative results of certain treatments in all possible demographic subpopulations will in turn bias the thinking of the clinical scientists who are planning the study. As a result, a serious selection bias will be introduced in the selection of patients and the corresponding treatment assignment.

As indicated in the previous chapter, one of the most critical questions often asked at the FDA advisory committee meeting is whether the patient population has been sufficiently characterized. It should be recognized that the patients in a clinical trial are only a sample with certain characterizations of demography and disease status defined by the inclusion and exclusion criteria of the protocol. The real question is whether the patients are a true

Table 2.3.2 Other Types of Bias

-
1. *Selection bias.* Physicians may recruit patients for clinical trials in ways that abuse the data.
 2. *Information bias.* The information patients provide to physicians (or others) is heavily tainted by their own beliefs and values.
 3. *Observer bias.* The objectivity of physicians or others who measure the magnitude of patient responses varies greatly, even in tests with objective endpoints.
 4. *Interviewer bias.* Interviewer bias is a well-known and obvious source of bias in clinical trials, particularly when interviews are used in measuring endpoints that determine the clinical trial's outcome.
 5. *Use of nonvalidated instruments.* The use of nonvalidated instruments is widespread in clinical trials.
 6. *Active control basis.* Biased higher cure rates for a new antifungal medicine compared with an active medicine rather than compared with a placebo.
-

Source: Spilker (1991).

representative sample of the targeted patient population. Ideally the patients in a clinical trial should be a representative sample randomly selected from the targeted patient population to which the results of the trial can be inferenced. However, in practice, most of clinical trials enrolled patients sequentially as long as they meet the inclusion and exclusion criteria. This practice is quite different from that of survey sampling or political poll for which samples can easily be selected at random using a telephone book or voter's registration list. For clinical trials it is almost impossible to achieve the ideal goal of random selection of representative sample from the targeted population. If the size of the study is quite large, for example, the GUSTO I trial (1993), valid inference and bias may not be issues for major efficacy and safety evaluation in a trial size of 41,000 patients. On the other hand, most clinical trials conducted by the pharmaceutical industry for registration are of small to moderate size. Bias will occur if care is not exercised in the selection of random samples from the targeted population. This bias is particularly crucial for a clinical development program because it will accumulate from phase I to phase III when adequate and well-controlled studies must produce substantial evidence for the approval of the drug.

In many clinical trials the enrollment is slow due to a seasoning of the disease or the geographical location. In such cases, the sponsors will open more study sites or enroll more patients at existing sites, whenever possible, in order to reach the required number of evaluable patients. For example, let us consider a phase II dose-ranging trial on basal cell carcinoma (a common skin cancer) conducted by a major pharmaceutical company. This study consisted of a placebo group and three treatment groups of low, medium, and high doses. The study called for a total of 200 patients with 50 patients per dose group. This study was a multicenter study with four study sites in the United States. Three sites were located in the south: San Diego, California, Phoenix, Arizona, and Houston, Texas, which are known to have high prevalence and incidence of skin cancers. The other site was Minneapolis, Minnesota. It turned out that the three sites in the south had no problem of enrollment but that the site at Minneapolis only enrolled a total of three patients. In order to finish the study on time, a decision was made not to open new sites but to enroll as many patients as possible at the three other sites. In order to meet the required sample size of 200 patients, this study ended up with one site of only three patients and three sites with 65 to 70 patients. As a result the validity of statistical inference based on the results from the study is doubtful due to the possible bias caused by the fact that one site in the north only enrolled three patients.

and the other three sites in the south enrolled most patients which may not constitute a representative sample of the targeted patient population.

Variability

As mentioned earlier, the reliability is referred to as the degree of the closeness (or precision) of the clinical results to the true value regarding the targeted patient population. The reliability of a clinical trial is an assessment of the precision of the clinical trial which measures the degree of the closeness of the clinical results to the true value. Therefore the reliability of a clinical trial reflects the ability to repeat or reproduce similar clinical outcomes in the targeted patient population to which the clinical trial is inferred. The higher precision a clinical trial has, the more likely the results will be reproducible. The precision of a clinical trial can be characterized by the variability of an estimated treatment effect based on some clinical endpoints used for clinical evaluation of the trial. In practice, sample size and the variability of the primary clinical endpoint play an important role in determining the precision and reproducibility of the clinical trial. The larger the sample size of the clinical trial is, the higher the precision and the more reliable the result will be. In clinical trials, however, the sample size is usually not large, and it cannot be increased indefinitely due to limited budget, resources, and often difficulty in patients recruitment. Indeed, the cost of achieving a desired precision can be extremely prohibitive. As an alternative, we can carefully define patient inclusion and exclusion criteria to reduce the variability of primary clinical endpoints and consequently reduce the cost. It is therefore critical that detailed procedures be implemented in the protocol to ensure that all the participating investigators keep to a clinical evaluation that is as homogeneous (less variable) as possible. To draw a reliable statistical inference on the efficacy and safety of the study medication, it is equally imperative to identify all sources of variations that may occur during the conduct of the trial. After these known sources of variation are identified, an appropriate statistical methodology based on the study design can be used to separate the known variabilities from any naturally inherent variability (or random error) for clinical evaluation of the study medication.

Note that for a clinical response consisting of several components, the variation of the clinical response may involve some known and unknown sources of variations. For example, a clinical response may include some continuous measurements such as systolic and diastolic blood pressures (mmHg) or direct bilirubin (mg/dL), ordinal categorical data such as NIH stroke scale (NIHSS) for quantification of neurologic deficit in the patients with acute ischemic stroke, or some binary data such as the cure of a patient infected with a certain bacteria by some antibiotic. All these clinical responses may be viewed as a sum of several components. One of the components may constitute the true unknown variation, and the others may consist of known sources of variations. These variations cause the variation of the clinical response.

Colton (1974) classifies sources of variation of quantitative clinical responses into three types: true biological and temporal variation and variability due to the measurement error. True biological variation is caused by the difference between subjects. In other words, factors such as age, gender, race, genotype, education status, smoking habit, sexual orientation, study center, and underlying disease characteristics at the baseline possibly can cause variation among subjects. True biological variation explains the differences among individuals. This source of variation is classified as a source of variation for the intersubject variability. The second type of the variation is the variability of clinical responses from the same

subject measured at different time points at which the status of the subject may change and vary. A well-known phenomenon of this type of variation is the circadian rhythms. For example, the systolic and diastolic blood pressures measured every 30 minutes by a 24-hour ambulatory device, premature ventricular contraction (PVC) as recorded on a 24-hour Holter monitor, or iron level measured every 8 hours over a 7-day period by either ICP or colorimetric methods all demonstrate the circadian rhythms of these clinical measurements. As a result, if the clinical responses exhibit temporal variation such as circadian rhythms, it is very important to eliminate this type of variation from comparison between treatments. Since the temporal variation reflects the fluctuation of clinical responses within the same subject, it is also known as a source of the intrasubject variability. Note that since the status for the cause of the temporal variation may be different treatments received by the same subject at different time points, this information may be used at the planning stage of clinical trials. Dose titration studies and crossover designs are typical examples utilizing temporal variation for comparison of treatments. As compared to parallel designs that involve both intersubject and intrasubject variabilities, dose titration studies and crossover designs utilizing intrasubject variabilities will provide better precision.

For assessment of efficacy and safety of study medication, a question of particular interest to clinical scientists is, Can similar clinical results be observed if the trial is to be repeatedly carried out under the same conditions? This concerns the variation of a clinical response due to the so-called measurement error. Measurement error is probably the most important variation that is difficult to detect and/or control. Measurement error induces the variation among the repeated measurements for the same clinical endpoint obtained from the same subject under the same environment. The possible causes of measurement error include observers such as clinicians or study nurses and laboratory errors caused by an instrument, the technician, or others. If a clinical endpoint produces an unacceptable measurement error, then the results cannot be reproduced under the same conditions. In this case the data have very little value in the clinical evaluation because they do not provide reliably the intended measurement. Since the variation due to measurement error is the variation of replication from the same subject, it can also be classified as a source of intra-subject variability.

Both true biological and temporal variations can be controlled by appropriate statistical designs, blocking, or stratification. Their impact on the precision of the estimates for treatment effects may be eliminated through adequate statistical analyses. In practice, the variation of a primary clinical endpoint due to measurement error is often used for the determination of sample size. Although variation due to measurement error cannot be eliminated completely, it can, however, be reduced tremendously by specifying standard procedures for measuring clinical endpoints in the protocol. Training can then take place, at the investigators' initiatives of the appropriate personnel who will be involved in the trial. Increasingly the training of clinicians, study nurses, and laboratory technicians is becoming important, since the data are gathered by sophisticated machinery, computers, or questionnaires. The clinical personnel must not only understand the rationale and newly developed technology but also be able to perform consistently throughout the study according to the procedures specified in the protocol. For example, the bone mineral density (BMD in g/cm³) of the spine is one of the primary and objective clinical responses for clinical evaluation of osteoporosis by certain interventions. A densitometer, which is a dual-energy X-ray absorptiometer, and its accompanying computer algorithm are used to determine bone mineral density. To ensure the reproducibility and consistency of BMD measurements, technicians are trained on the densitometer using an anthropomorphic spine phantom with a User's Manual provided by the manufacturer. In

addition the spine phantom be used daily for calibration to lessen the drift effect of the instrument. The regular quality control evaluation should be maintained by the manufacturer. Another source of variation for BMD measurements is the position of the patients. In order to reduce this variation, the positioning of the patient at the various visits should be as close as possible to that at his/her baseline visit. Since the measurement error associated with BMD determination is mainly with the performance of technician, it is preferable that one technician carry out the measurements through the entire study.

Another example for possible reduction of measurement error by training is the NIH Stroke Scale (NIHSS) which was developed by the U.S. National Institute of Neurologic Disorder and Stroke (NINDS) from the original scale devised at the University of Cincinnati. Technically, it is a rating scale for quantifying a neurological deficit by a total of 42 points in 11 categories such as given in Table 2.3.3. In practice, it is a rating scale based on the subjective judgment of qualified neurologists and emergency physicians. This scale has recently been used in several clinical trials as the primary outcome measure for efficacy assessment in various thrombolytic agents in patients with acute ischemic stroke (The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group, 1995). In these multicenter clinical trials, in order to maintain double-binding, different clinicians were asked to perform the baseline and subsequent NIHSS. Standardization of the procedure for rating NIHSS in the protocol by the clinicians within the same center as well as by investigators at different centers is considered crucial in order to reduce measurement error and produce reliable, consistent, and reproducible results. As a result, video training was selected to train and certify investigators who perform NIHSS so that the measurement error in those clinical trials could be reduced.

In summary, to have an accurate and reliable assessment of the true efficacy and safety of a study medication, it is important to avoid bias and to minimize the variability of the primary clinical endpoint whenever possible. Figure 2.3.1 shows the impact of bias and variability in tackling the truth. As can be seen from the figure, the ideal situation is to have an unbiased estimate with no variability. If the estimate has a nonnegligible bias, the truth is compromised even for the smallest variability. If the estimate is biased with a lot of variability, the assessment for the treatment effect must be discarded.

2.4 CONFOUNDING AND INTERACTION

In clinical trials, confounding and interaction effects are the most common distortions in the evaluation of medication. *Confounding effects* are contributed by various factors such as race and gender that cannot be separated by the design under study; an *interaction effect* between factors is a joint effect with one or more contributing factors (Chow and Liu, 1995a). Confounding and interaction are important considerations in clinical trials. For example, when confounding effects are observed, we cannot assess the treatment effect because it is contaminated by other effects. On the other hand, when interactions among factors are observed, the treatment must be carefully evaluated for those effects.

Confounding

In clinical trials, there are many sources of variation that have an impact on the primary clinical endpoints for evaluation relating to a certain new regimen or intervention. If some of these variations are not identified and properly controlled, they can become

Table 2.3.3 Summary of the National Institutes of Health Stroke Scale

Item	Name	Response
1A	Level of consciousness	0 = Alert 2 = Not alert, obtunded 3 = Unresponsive
1B	Questions	0 = Answers both correctly 1 = Answers one correctly 2 = Answers neither correctly
1C	Commands	0 = Performs both tasks correctly 1 = Performs one task correctly 2 = Performs neither correctly
2	Gaze	0 = Normal 1 = Partial gaze palsy 2 = Total gaze palsy
3	Visual fields	0 = No visual loss 1 = Partial hemianopsia 2 = Complete hemianopsia 3 = Bilateral hemianopsia
4	Facial palsy	0 = Normal 1 = Minor paralysis 2 = Partial paralysis 3 = Complete paralysis
5	Motor arm a. Left b. Right	0 = No drift 1 = Drift before 10 seconds 2 = Fall before 10 seconds 3 = No effort against gravity 4 = No movement
6	Motor leg a. Left b. Right	0 = No drift 1 = Drift before 5 seconds 2 = Fall before 5 seconds 3 = No effort against gravity 4 = No movement
7	Ataxia	0 = Absent 1 = One limb 2 = Two limbs
8	Sensory	0 = Normal 1 = Mild loss 2 = Severe loss
9	Language	0 = Normal 1 = Mild aphasia 2 = Severe aphasia 3 = Mute or global aphasia
10	Dysarthria	0 = Normal 1 = Mild 2 = Severe
11	Extinction/inattention	0 = Normal 1 = Mild 2 = Severe

Source: Lyden et al. (1994).

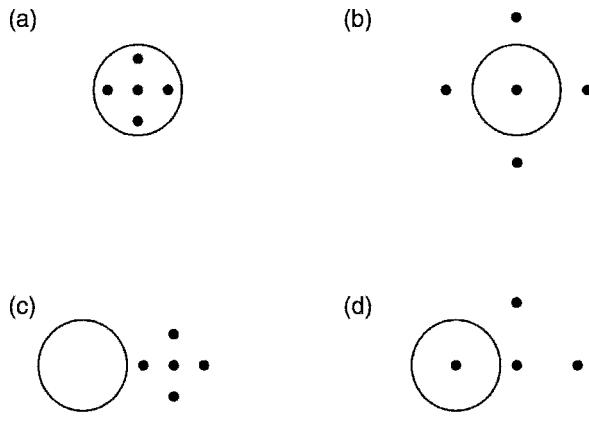


Figure 2.3.1 Accuracy and precision for assessment of treatment effect. (a) Unbiased and least variability, (b) unbiased and large variability, (c) biased and least variability, and (d) biased and large variability.

mixed in with the treatment effect that the trial is designed to demonstrate. Then the treatment effect is said to be confounded by effects due to these variations. To provide a better understanding, consider the following example. Suppose that last winter Dr. Smith noticed that the temperature in the emergency room was relatively low and caused some discomfort among medical personnel and patients. Dr. Smith suspected that the heating system might not be functioning properly and called on the hospital to improve it. As a result, the temperature of the emergency room is at a comfortable level this winter. However, this winter is not as cold as last winter. Therefore, it is not clear whether the improvement in the emergency room temperature was due to the improvement in the heating system or the effect of a warmer winter. In fact, the effect due to the improvement of the heating system and that due to a warmer winter are confounded and cannot be separated from each other. In clinical trials, there are many subtle, unrecognizable, and seemingly innocent confounding factors that can cause ruinous results of clinical trials. Moses (1992) gives the example of the devastating result in the confounder being the personal choice of a patient. The example concerns a polio-vaccine trial that was conducted on two million children worldwide to investigate the effect of Salk poliomyelitis vaccine. This trial reported that the incidence rate of polio was lower in the children whose parents refused injection than those who received placebo after their parent gave permission (Meier, 1989). After an exhaustive examination of the data, it was found that susceptibility to poliomyelitis was related to the differences between the families who gave the permission and those who did not.

Sometimes, confounding factors are inherent in the design of the studies. For example, dose titration studies in escalating levels are often used to investigate the dose-response relationship of the antihypertensive agents during phase II stage of clinical development. For a typical dose titration study, after a washout period during which previous medication stops and the placebo is prescribed, N subjects start at the lowest dose for a prespecified time interval. At the end of the interval, each patient is evaluated as a responder to the treatment or a non-responder according to some criteria prespecified in the protocol. In a titration study, a subject will continue to receive the next higher dose if he or she fails, at the

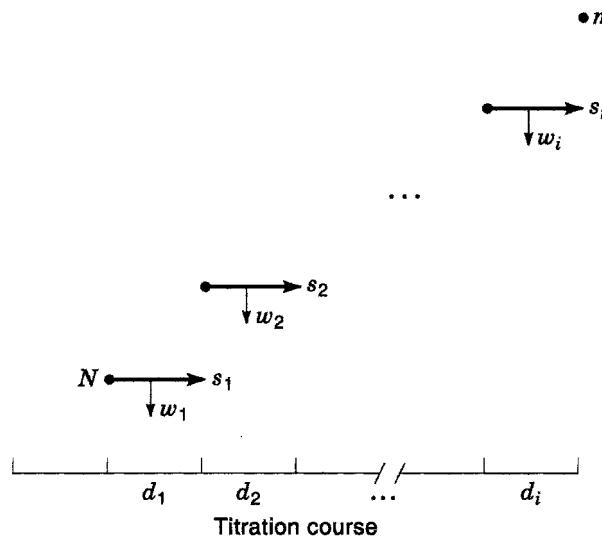


Figure 2.4.1 Graphical display of a titration trial. d_i , the i th dose level; s_i , the number of subjects who responded at the i th dose; w_i , the number of subjects who withdrew at the i th dose; and m , the number of subjects who completed the study without a response. (Source: Shih, Gould, and Hwang, 1989.)

current level, to meet some objective physiological criteria such as reduction of diastolic blood pressure by a prespecified amount and has not experienced any unacceptable adverse experience. Figure 2.4.1 provides a graphical presentation of a typical titration study (Shih, Gould, and Hwang, 1989). Dose titration studies are quite popular among clinicians because they mimic real clinical practice in the care of patients. The major problem with this typical design for a dose titration study is that the dose-response relationship is often confounded with time course and the unavoidable carryover effects from the previous dose levels which cannot be estimated and eliminated. One can always argue that the relationship found in a dose titration study is not due to the dose but to the time. Statistical methods for binary data from dose titration studies have been suggested under some rather strong assumptions (e.g., see Chuang, 1987; Shih, Gould, and Hwang, 1989). Due to the fact that the dose level is confounded with time, estimation of the dose-response relationship based on continuous data has not yet been resolved. Another type of design that can induce confounding problems when it is conducted inappropriately is the crossover design. For a standard 2×2 crossover design, each subject is randomly assigned to one of the two sequences. In sequence 1, subjects receive the reference (or control) treatment at the first dosing period and the test treatment at the second dosing period after a washout period of sufficient length. The order of treatments is reversed for the subjects in sequence 2. The issues in analysis of the data from a 2×2 crossover design is twofold. First, unbiased estimates of treatment effect cannot be obtained from the data of both periods in the presence of a nonzero carryover effect. The second problem is that the carryover effect is confounded with sequence effect and treatment-by-period interaction. In absence of a significant sequence effect, however, the treatment effect can be estimated unbiasedly from the data of both periods. In practice, it is not clear whether an observed statistically significant sequence effect (or carryover effect) is a true sequence effect (or carryover effect). As

a result this remains a major drawback of the standard 2×2 crossover design, since the primary interest is to estimate a treatment effect that is still an issue in the presence of a significant nuisance parameter. The sequence and carryover effects, however, are not confounded to each other in higher-order crossover designs that compare two treatments and can provide unbiased estimation of treatment effect in the presence of a significant carry-over effect (Chow and Liu, 1992, 2000).

Bailar (1992) provided another example of subtle and unrecognizable confounding factors. In the same issue of *New England Journal of Medicine*, Wilson et al. (1985) and Stampfer et al. (1985) both reported the results on the incidence of cardiovascular diseases in postmenopausal women who had been taking hormones compared to those who had not. Their conclusions, however, were quite different. One reported that the incidence rate of cardiovascular disease among the women taking hormones was twice that in the control group, while the other reported a totally opposite conclusion in which the incidence of the experimental group was only half that of women who were not taking hormones. Although these trials were not randomized studies, both studies were well planned and conducted. Both studies had carefully considered the differences in known risk factors between the two groups in each study. As a result, the puzzling difference in the two studies may be due to some subtle confounding factors such as the dose of hormones, study populations, research methods, and other related causes. This example indicates that it is imperative to identify and take into account all confounding factors for the two adequate, well-controlled studies that are required for demonstration of effectiveness and safety of the study medication.

In clinical trials, it is not uncommon for some subjects not to follow instructions in taking the prescribed dose at the scheduled time as specified in the protocol. If the treatment effect is related to (or confounded with) patients' compliance, any estimates of the treatment effect are biased unless there is a placebo group in which the differences in treatment effects between subjects with good compliance and poor compliance can be estimated. As a result, interpretation and extrapolation of the findings are inappropriate. In practice, it is very difficult to identify compliers and noncompliers and to quantify the relationship between treatment and compliance. On the other hand, subject withdrawals or dropouts from clinical trials are the ultimate examples of noncompliance. There are several possible reasons for dropouts. For example, a subject with severe disease did not improve and hence dropped out from the study. The estimate of treatment effect will be biased in favor of a false positive efficacy, if the subjects with mild disease remain and improve. On the other hand, subjects will withdraw from a study if their conditions improve, and those who did not improve will remain until the scheduled termination of a study. The estimation of efficacy will then be biased and hence indicate a false negative efficacy. Noncompliance and subject dropouts are only two of the many confounding factors that can occur in many aspects of clinical trials. If there is an unequal proportion of the subjects who withdraw from the study or comply to the dosing regimen among different treatment groups, it is very important to perform an analysis on these two groups of subjects to determine whether confounded factors exist and the direction of possible bias. In addition every effort must be made to continue subsequent evaluation of withdrawals in primary clinical endpoints such as survival or any serious adverse events. For analyses of data with noncompliance or withdrawals, it is suggested that an *intention-to-treat* analysis be performed. An intention-to-treat analysis includes all available data based on all randomized subjects with the degree of compliance or reasons for withdrawal as possible covariates.

Interaction

The objective of a statistical interaction investigation is to conclude whether the joint contribution of two or more factors is the same as the sum of the contributions from each factor when considered alone. The factors may be different drugs, different doses of two drugs, or some stratification variables such as severity of underlying disease, gender, or other important covariates. To illustrate the concept of statistical interaction, we consider the Second International Study of Infarct Survival (ISIS-2, 1988). This study employed a 2×2 factorial design (two factor with two levels at each factor) to study the effect of streptokinase and aspirin in the reduction of vascular mortality in patients with suspected acute myocardial infarction. The two factors are one-hour intravenous infusion of 1.5 MU of streptokinase and one month of 150 mg per day enteric-coated aspirin. The two levels for each factor are either active treatment and their respective placebo infusion or tablets. A total of 17,187 patients were enrolled in this study. The numbers of the patients randomized to each arm is illustrated in Table 2.4.1. The key efficacy endpoint is the cumulative vascular mortality within 35 days after randomization. Table 2.4.2 provides the cumulative vascular mortality for each of the four arms as well as those for streptokinase and aspirin alone. From Table 2.4.2 the mortality of streptokinase group is about 9.2%, with the corresponding placebo mortality being 12.0%. The improvement in mortality rate attributed to streptokinase is 2.8% ($12.0\% - 9.2\%$). This is referred to as the main effect of streptokinase. Similarly the main effect of aspirin tablets can also be estimated from Table 2.4.2 as 2.4% ($11.8\% - 9.4\%$). The left two panels of Figure 2.4.2 give the cumulative vascular mortalities of main effects for both streptokinase and placebo. The right panel of Figure 2.4.2 provides mortality for combination of streptokinase and aspirin against that of both placebos. From either Table 2.4.2 or figure 2.4.2, the joint contribution of both

Table 2.4.1 Treatment of ISIS-2 with Number of Patients Randomized

		IV Infusion of Streptokinase		
Aspirin		Active	Placebo	Total
Active		4292	4295	8,587
Placebo		4300	4300	8,600
Total		8592	8595	17,187

Source: ISIS-2 (1988).

Table 2.4.2 Cumulative Vascular Mortality in Days 0–35 of ISIS-2

		IV Infusion of Streptokinase		
Aspirin		Active	Placebo	Total
Active		8.0%	10.7%	9.4%
Placebo		10.4%	13.2%	11.8%
Total		9.2%	12.0%	

Source: ISIS-2 (1988).

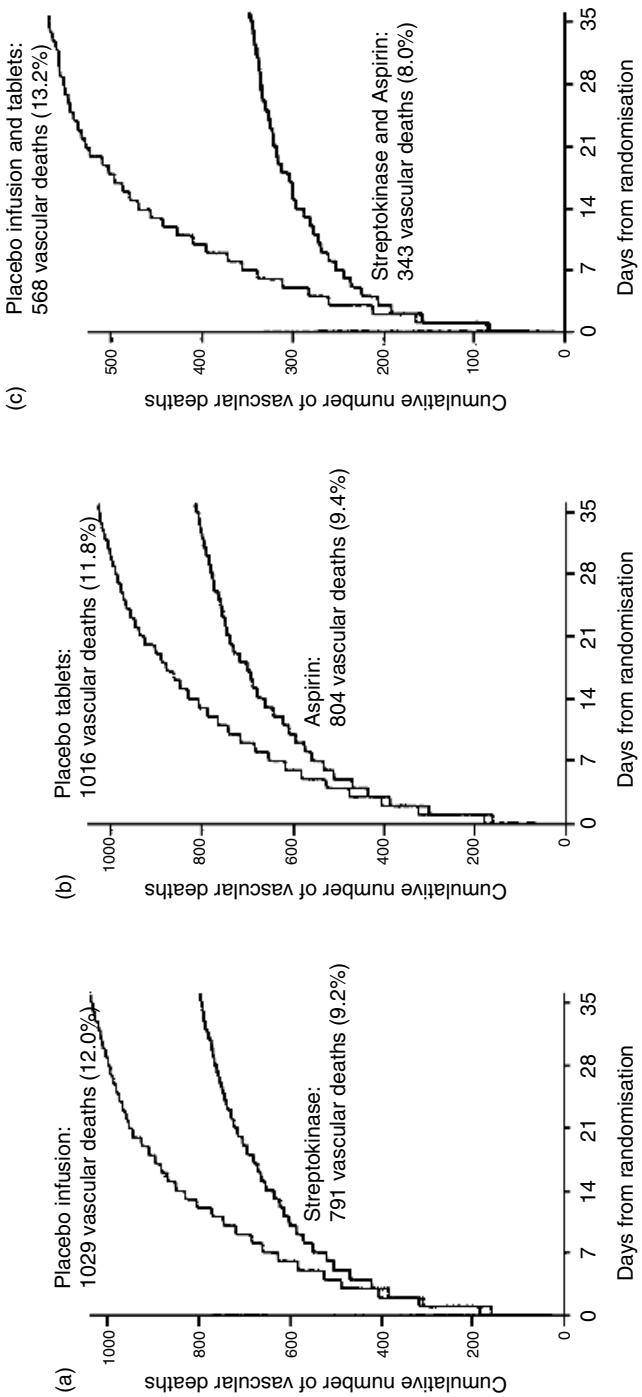


Figure 2.4.2 Cumulative vascular mortality in days 0–35. (Source: ISIS-2, 1988.)

streptokinase and aspirin in improvement in mortality is 5.2% ($13.2\% - 8.0\%$) which is exactly equal to the contribution in mortality by streptokinase (2.8%) plus that by aspirin (2.4%). This is a typical example that no interaction exists between streptokinase and aspirin because the reduction in mortality by joint administration of both streptokinase and aspirin can be expected as the sum of reduction in mortality attributed to each antithrombolytic agent when administrated alone. In other words, the difference between the two levels in one factor does not depend on the level of the other factor. For example, the difference in vascular mortality between streptokinase and placebo for the patients taking aspirin tablets is 2.7% ($10.7\% - 8.0\%$). A similar difference of 2.8% is observed between streptokinase (10.4%) and placebo (13.2%) for the patients taking placebo tablets. Therefore the reduction in mortality attributed to streptokinase is homogeneous for the two levels of aspirin tablets. As a result, there is no interaction between streptokinase infusion and aspirin tablets. This phenomenon is also observed in Figure 2.4.3.

The ISIS-2 trial provides an example of an investigation of interaction between two treatments. However, in the clinical trial it is common to check interaction between treatment and other important prognostic and stratification factors. For example, almost all adequate well-controlled studies for the establishment of effectiveness and safety for approval of pharmaceutical agents are multicenter studies. For multicenter trials, the FDA requires that the treatment-by-center interaction be examined to evaluate whether the treatment effect is consistent across all centers.

One of the objectives of the National Institute of Neurological Disorders and Stroke rt-PA Stroke Study is to investigate whether the improvement of neurological deficit upon

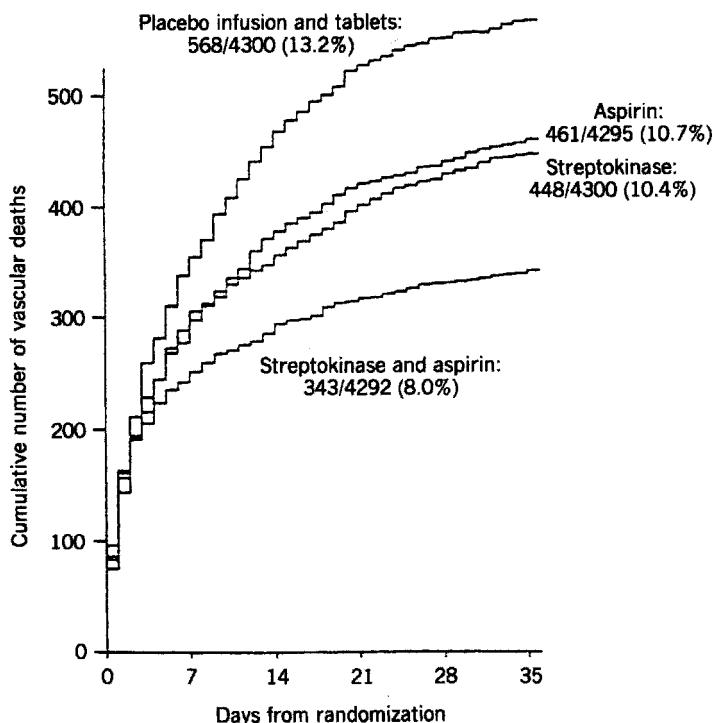


Figure 2.4.3 Cumulative vascular mortality in days 0–35 by treatment groups. (Source: ISIS-2, 1988.)

administration of intravenous recombinant t-PA over the placebo group is consistent over the time to receiving treatment after the onset of stroke as stratified from 0 to 90 minutes and from 90 to 180 minutes. The results based on NIHSS are reproduced in Table 2.4.3. It can be seen that the difference in NIHSS between t-PA and placebo (i.e., treatment effect) is homogeneous between the two time intervals. Consequently, no interaction exists between the treatment and time to treatment.

When the difference among levels of one factor is not the same at different levels of other factors, then it is said that interaction exist between these two factors. In general, interactions can be classified as quantitative or qualitative (Gail and Simon, 1985). A quantitative interaction is the one for which the magnitude of the treatment effect is not the same across the levels of other factors but the direction of the treatment remains the same for all levels of other factors. A qualitative interaction is the interaction in which the direction of the treatment effect changes in some levels of other factors. To provide a better understanding, we consider the following hypothetical example of the treatment of an irreversible inhibitor of steroid aromatase in the patients with benign prostatic hyperplasia. Suppose that one of the objectives of the trial is to investigate whether improvement of peak urinary flow rate (mL/sec) of the treatment over placebo is the same for patients with an American Urinary Association (AUA) symptom score between 8 and 19 inclusively and those with AUA score greater than 19. There are two factors, each with two levels. One can display the four treatment-by-symptom means in a figure for visual inspection of possible interaction. The vertical axis is the mean change from baseline in peak urinary flow rate (mL/sec). The two levels of treatment can be represented on the horizontal axis. The mean peak urinary flow rate of all levels of AUA symptom score (other factor) then can be plotted at each level of treatment on the horizontal axis, and the means of the same levels of AUA symptom score are connected over the two levels of treatment. Panel A of Figure 2.4.4 exhibits a pattern of no interaction in which the two lines are parallel and the distance between the two lines is the same at all levels of treatment. Panels B and C demonstrate a possible quantitative interaction between the treatment and AUA symptom score because the two lines do not cross and treatment effect does not change its direction, though the distance between the lines is not the same. Both panels B and C indicate that the treatment is more effective than placebo for increasing peak flow rate. Panel B shows that treatment induces a better improvement in peak urinary flow rate in patients with an AUA symptom score greater than 19 than those between 8 and 19, while the opposite observation is seen in panel C. In panel D a positive difference in peak urinary flow rate between treatment and placebo is observed in patients with an AUA symptom score greater than 19, while the difference is negative in patients with an AUA symptom score

Table 2.4.3 Mean NIHSS Score of the National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group

Time to Treatment after Stroke Onset	t-PA	Placebo
0–90 minutes	$N = 157$ 9(2–17)	$N = 145$ 12(6–18)
0–180 minutes	$N = 155$ 8(3–19)	$N = 167$ 13(7–19)

Source: National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group (1995).

Note: Number in the parentheses are ranges.

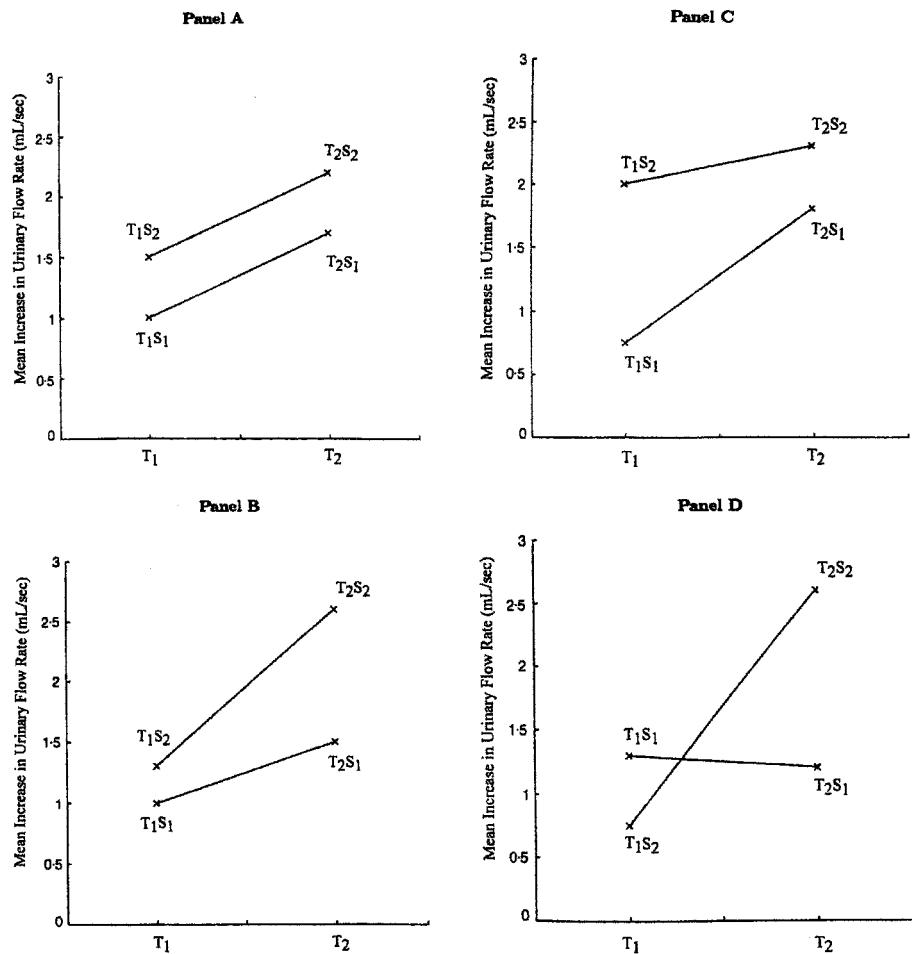


Figure 2.4.4 Graphical presentation of two-factor interaction.

between 8 and 19. Panel *D* shows a possible qualitative interaction between treatment and symptom score. Since the two lines cross each other in panel *D*, a qualitative interaction of this kind is also called a crossover interaction.

Interaction can be used to investigate whether the effectiveness of treatment is homogeneous across groups of patients with different characteristics. It therefore is important in the interpretation and inference of the trial results. Gail and Simon (1985) provided an graphical illustration of different interactions for two subgroups of patients, which is reproduced in Figure 2.4.5. The true differences in efficacy between two treatments in subgroups 1 and 2 are δ_1 and δ_2 . The 45° line is the line where $\delta_1 = \delta_2$. It therefore represents the line of no interaction. In the unshaded areas (the first and third quadrants) δ_1 and δ_2 are either both positive or both negative. Any point in the unshaded area except for those on the 45° line represents a quantitative interaction. The two shaded areas (the second and fourth quadrants) consist of points for qualitative (or crossover) interaction.

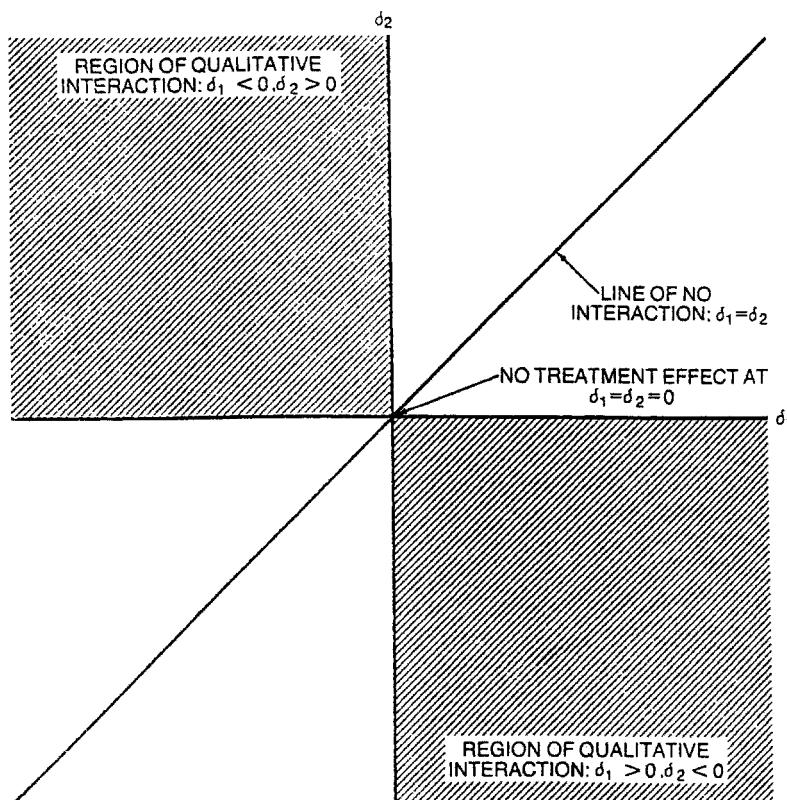


Figure 2.4.5 Interaction of two subgroups. (Source: Gail and Simon, 1985.)

2.5 DESCRIPTIVE AND INFERRENTIAL STATISTICS

Dietrich and Kearns (1986) divided statistics into two broad areas, namely descriptive and inferential statistics. Descriptive statistics is the science of summarizing or describing data, while inferential statistics is the science of interpreting data in order to make estimates, hypotheses testing, predictions, or decisions from the samples to the targeted population.

In clinical trials, data are usually collected through case report forms which are designed to capture clinical information from the studies. The information on the case report forms is then entered into the database. The raw database is always messy, though it does contain valuable clinical information from the study. In practice, it is often of interest to summarize the raw database by a graphical presentation (e.g., a data plot) or by descriptive (or summary) statistics. Descriptive statistics are simple sample statistics such as means and standard deviations (or standard errors) of clinical variables or endpoints. Note that the standard deviation describes the variability of a distribution, either a population distribution or a sample distribution, whereas the standard error is the variability of a sample statistic (e.g., sample mean or sample variance). Descriptive statistics are often used to describe the targeted population before and after the study. For example, at baseline, descriptive statistics are often employed to describe the comparability between treatment

groups. After the completion of the study, descriptive statistics are useful tools to reveal possible clinical differences (or effects) or trends of study drugs. As an example, Table 2.5.1 provides a partial listing of individual patient demographics and baseline characteristics from a study comparing the effects of captopril and enalapril on quality of life in the older hypertensive patients (Testa et al., 1993). As can be seen from Table 2.5.1, although as a whole, the patient listing gives a detailed description of the characteristics for individual patients, it does not provide much summary information regarding the study population. In addition descriptive statistics for demographic and baseline information describe not only the characteristics of the study population but also the comparability between treatment groups (see Table 2.5.2). In addition, for descriptive purposes, Table 2.5.3 groups patients into low, medium, and high categories according to the ranking of their scores on the baseline quality of life scale. It can be seen that there is a potential difference in treatment effect among the three groups with regard to the change from baseline on the quality of life. These differences were confirmed to be statistically significant by valid statistical tests. Therefore a preliminary investigation of descriptive statistics of primary clinical endpoints may reveal a potential drug effect.

When we observe some potential differences (effects) or trends, it is necessary to further confirm with certain assurance that the differences (effects) or trends indeed exist and are not due to chance alone. For this purpose it is necessary to provide inferential statistics for the observed differences (effects) or trends. Inferential statistics such as confidence intervals and hypotheses testing are often performed to provide statistical inference on the possible differences (effects) or trends that can be detected based on descriptive statistics. For the rest of this section, we will focus on confidence intervals (or interval estimates). Hypotheses testing will be discussed in more detail in the following section.

Clinical endpoints are often used to assess the efficacy and safety of drug products. For example, diastolic blood pressure is one of the primary clinical endpoints for the study of ACE inhibitor agents in the treatment of hypertensive patients. The purpose of the diastolic blood pressure for hypertensive patients is to compare their average diastolic blood pressure with the norm for ordinary health subjects. However, the average diastolic blood pressure for the hypertensive patients is unknown. We will need to estimate the average diastolic blood pressure based on the observed diastolic pressures obtained from the hypertensive patients. The observed diastolic blood pressures and the average of these diastolic blood pressures are the sample and sample mean of the study. The sample mean is an estimate of the unknown population average diastolic blood pressure. Point estimates may not be of practical use. For example, suppose that the sample mean is 98 mmHg. It is then important to know whether the population average for the hypertensive patients could reasonably be 90 mmHg given that the sample average turned out to be 98 mmHg. This kind of information depends on the knowledge of the standard error, not merely of the point estimate itself.

The observed diastolic blood pressures are usually scattered around the sample mean. Based on these observed diastolic blood pressures, the standard error of the sample mean of the observed diastolic blood pressures can be obtained. If the distribution of the diastolic blood pressure appears to be a bell shaped and the sample size is of moderate size, then there is about 95% chance that the unknown average diastolic blood pressure of the targeted population will fall within the area between approximate two (i.e., 1.96) standard errors below and above the sample mean. The lower and upper limits of the area constitute an interval estimate for the unknown population average diastolic blood pressure. An interval estimate is usually referred to as a confidence interval with a desired confidence level,

Table 2.5.1 Partial Data of Capoten Quality of Life Study

Patient	Race	Age (years)	Height (inches)	Weight (pounds)	Heart Rate (per minute)	Systolic BP (mmHg)	Diastolic BP (mmHg)	Alcohol Consumption	Tobacco Consumption
C01-003	Caucasian	69	70	195	72	146	94	No	No
C01-004	Caucasian	69	70	188	72	170	94	Yes	Yes
C01-006	Caucasian	62	76	231.5	84	158	91	No	No
C01-008	Caucasian	56	72	244	76	140	97	No	No
C01-010	Caucasian	55	75	258	60	139	100	Yes	No
C01-012	Caucasian	58	74	191	86	138	95	Yes	No
C02-004	Black	66	70	220	64	141	99	No	No
C02-006	Black	55	65	244.5	64	171	109	No	No
C02-008	Black	61	69	281	76	173	104	No	No
C02-009	Black	61	68	190.5	60	150	91	Yes	No
C02-013	Black	71	74	193	88	140	97	No	Yes
C03-001	Caucasian	55	74	303	76	151	103	Yes	No
C03-004	Caucasian	65	71	243	78	156	91	No	No
C03-005	Caucasian	55	69	178	100	150	91	Yes	No
C03-006	Caucasian	59	65	174	64	157	101	Yes	Yes
C03-010	Caucasian	74	67	171	80	188	109	No	No
C03-012	Caucasian	65	65	150	58	169	99	Yes	No
C04-003	Caucasian	64	72	194	96	161	99	Yes	No
C04-005	Black	59	69	201	76	179	96	No	No
C04-009	Caucasian	58	78	334	80	159	114	No	No

Table 2.5.2 Demographic, Clinical, and Quality of Life Variables at the Baseline

Variable	Captopril (N = 192)	Enalapril (N = 187)
Demographic		
Age (yr)	64.2 ± 5.5	64.6 ± 6.4
Education (%)		
No high school	9	7
Some high school	36	30
Some college	49	58
Postgraduate degree	6	5
Income (%)		
<\$15,000	15	12
\$15,000–40,999	41	47
\$41,000–80,000	33	35
>\$80,000	11	6
Percent married	87	88
Occupational status (%)		
Employed full-time	40	36
Employed part-time	14	15
Retired	43	49
Unemployed	3	1
Race (%)		
White	84	82
Black	14	18
Other	3	1
Clinical		
Weight (lb)	197.8 ± 36.4	198.7 ± 37.9
Body-mass index	28.8 ± 4.7	28.6 ± 5.0
Blood pressure (mmHg)		
Systolic	155.0 ± 14.8	154.6 ± 15.8
Diastolic	97.3 ± 5.8	97.3 ± 5.9
Previous antihypertensive therapy (%)	89	87
Quality-of-life scales		
Psychological well-being	462 ± 78	452 ± 80
Psychological distress	526 ± 64	521 ± 59
General perceived health	493 ± 77	495 ± 67
Well-being at work or in daily routine	485 ± 61	480 ± 62
Sexual-symptom distress	518 ± 144	503 ± 162
Distress and stress indexes		
Side effects and symptoms distress	24 ± 40	25 ± 35
Life events	30 ± 40	28 ± 40
Stress	274 ± 144	296 ± 142

Source: Testa et al. (1993).

such as 95%. Unlike a point estimate, a confidence interval provides a whole interval as an estimate for a population parameter instead of just a single value. A 95% confidence interval is a random interval that is calculated according to a certain procedure that would produce a different interval for each sample upon repeated sampling from the population, and 95% of these intervals would contain the unknown fixed population parameter. A 95%

Table 2.5.3 Changes from the Baseline to End Point in Quality of Life According to Scores on the Quality of Life Scale at the Baseline

Scale	Baseline Scores for Randomized Patients			Captopril (N = 184)			Enalapril (N = 178)		
	Low	Medium	High	Low	Medium	High	Low	Medium	High
				(N = 53)	(N = 60)	(N = 71)	(N = 60)	(N = 65)	(N = 53)
General perceived health	420	506	552	+21.0 ± 6.6	-2.7 ± 6.3	-1.6 ± 3.2	+1.7 ± 7.3	-10.4 ± 5.3	-15.8 ± 6.1
Psychological well-being	374	469	524	+19.3 ± 8.8	-1.4 ± 7.3	+8.2 ± 3.0	+17.7 ± 9.0	+1.3 ± 6.1	-7.9 ± 6.0
Psychological distress	457	533	577	+19.7 ± 6.8	-6.2 ± 5.9	-3.5 ± 2.6	+7.2 ± 6.5	-0.9 ± 5.1	-10.8 ± 4.5
Overall quality of life	427	502	545	+18.1 ± 5.3	-6.8 ± 5.4	-0.5 ± 2.4	+5.9 ± 6.1	-4.3 ± 4.5	-10.7 ± 4.6

Source: Testa et al. (1993).

Note: Mean ± SD change from baseline score.

confidence interval is not an interval that will contain 95% of the sample averages that would be obtained on repeating the sampling procedure, nor is a particular 95% confidence interval in which the population average will fall 95% of the times. It should be noted that the population average is an unknown constant and does not vary while a confidence interval is random. It is either in the confidence interval or not.

A classical confidence interval for the population average of a clinical variable is symmetric about the observed sample mean of the observed responses of the clinical variable. This classical confidence interval is sometimes called the shortest confidence interval because its width is the shortest among all of the confidence intervals of the same confidence level by other statistical procedures. In some situations, it may be of interest to obtain a symmetric confidence with respect to a fixed number. For example, in bioequivalence trials it is of interest to obtain a confidence interval for the difference in a pharmacokinetic parameter such as area under the blood or plasma concentration time curve (AUC) between the test and reference drug product. If the 90% confidence interval falls within $\pm 20\%$ of the average of the reference product, then we conclude that the test product is bioequivalent to the reference product (e.g., Chow and Liu, 2000). Since the limits are $\pm 20\%$, which is symmetric about 0%, Westlake (1976) proposed the idea to consider a symmetric confidence interval with respect to 0 rather than the shortest confidence interval symmetric about the observed difference in the sample means. Note that as indicated in Chow and Liu (2000), the most common criticisms of Westlake's symmetric confidence interval are that it has shifted away from the direction in which the sample difference was observed and that the tail probabilities associated with Westlake's symmetric confidence interval are not symmetric. As a result Westlake's symmetric confidence interval moves from a two-sided to a one-sided approach as the true difference and the random error increase.

The confidence level is the degree of certainty that the interval actually contains the unknown population parameter value. It provides the degree of assurance or confidence that the statement regarding the population parameter is correct. The more certainty we want, the wider the interval will have to be. A very wide interval estimate may not be of practical use because it fails to identify the population parameter closely. In practice, it is more usual to use 90%, 95%, or 99% as confidence levels. Table 2.5.4 summarizes the multiple of standard errors that are needed for confidence levels of 68, 95, and 99. Therefore we will have 68%, 95%, and more than 99% confidence that the population parameter will fall within one, two, and three standard errors of the observed value, respectively.

Table 2.5.4 Confidence Levels with Various Standard Errors

Standard Errors	Confidence Level
0.5	0.3830
0.675	0.5000
1.0	0.6826
1.5	0.8664
1.96	0.9500
2.0	0.9544
2.5	0.9876
3.0	0.9974
4.0	1.0000

When we claim that a drug product is effective and safe with 95% assurance, it is expected that we will observe consistent significant results 95% of times if the clinical trial were repeatedly carried out with the same protocol. However, current FDA regulation only requires two adequate well-controlled clinical trials be conducted to provide substantial evidence for efficacy and safety. It is therefore of interest to estimate the probability that the drug is effective and safe based on clinical results obtained from the two adequate well-controlled trials.

2.6 HYPOTHESES TESTING AND *p*-VALUES

In clinical trials a hypothesis is a postulation, assumption, or statement that is made about the population regarding the efficacy, safety, or other pharmacoeconomics outcomes (e.g., quality of life) of a drug product under study. This statement or hypothesis is usually a scientific question that needs to be investigated. A clinical trial is often designed to address the question by translating it into specific study objective(s). Once the study objective(s) has been carefully selected and defined, a random sample can be drawn through an appropriate study design to evaluate the hypothesis about the drug product. For example, a scientific question regarding a drug product, say drug *A*, of interest could be either (1) Is the mortality reduced by drug *A*? or (2) Is drug *A* superior to drug *B* in treating hypertension? The hypothesis to be questioned is usually referred to as the *null hypothesis*, denoted by H_0 . The hypothesis that the investigator wishes to establish is called the *alternative hypothesis*, denoted by H_a . In practice, we attempt to gain support for the alternative hypothesis by producing evidence to show that the null hypothesis is false. For the questions regarding drug *A* described above, the null hypotheses are that (1) there is no difference between drug *A* and the placebo in the reduction of mortality and (2) there is no difference between drug *A* and drug *B* in treating hypertension, respectively. The alternative hypotheses are that (1) drug *A* reduces the mortality and (2) drug *A* is superior to drug *B* in treating hypertension, respectively. These scientific questions or hypotheses to be tested can then be translated into specific study objectives as to compare (1) the efficacy of drug *A* with no therapy in the prevention of reinfarction of (2) the efficacy of drug *A* with that of drug *B* in reducing blood pressure in elderly patients, respectively.

Chow and Liu (2000) recommended the following steps be taken to perform a hypothesis testing:

1. Choose the null hypothesis that is to be questioned.
2. Choose an alternative hypothesis that is of particular interest to the investigators.
3. Select a test statistic, and define the rejection region (or a rule) for decision making about when to reject the null hypothesis and when not to reject it.
4. Draw a random sample by conducting a clinical trial.
5. Calculate the test statistic and its corresponding *p*-value.
6. Make conclusion according to the predetermined rule specified in step 3.

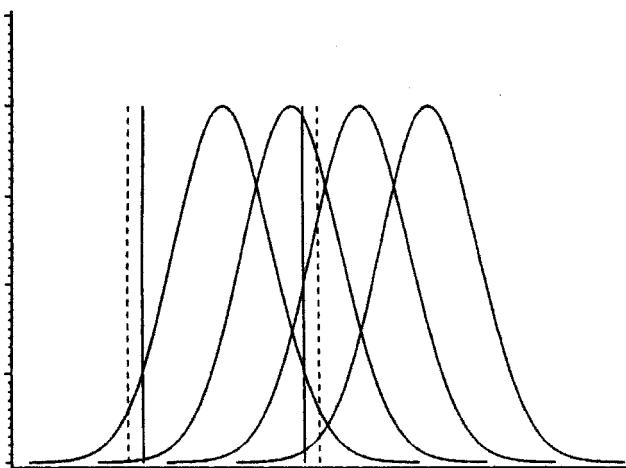
When performing a hypotheses testing, basically two kinds of errors occur. If the null hypothesis is rejected when it is true, then a type I error has occurred. For example, a type I error has occurred if we claim that drug *A* reduces the mortality when in fact there is no difference between drug *A* and the placebo in the reduction of mortality. The probability of

Table 2.6.1 Relationship Between Type I and Type II Errors

		If H_0 is	
		True	False
When	Fail to reject	No error	Type II error
	Reject	Type I error	No error

committing type I error is known as the level of significance. It is usually denoted by α . In practice, α represents the consumer's risk which is often chosen to be 5%. On the other hand, if the null hypothesis is not rejected when it is false, then a type II error has been made. For example, we have made a type II error if we claim that there is no difference between drug *A* and the placebo in the reduction of mortality when in fact drug *A* does reduce the mortality. The probability of committing type II error, denoted by β , is sometimes referred to as the producer's risk. In practice, $1 - \beta$ is known as the power of the test, which represents the probability of correctly rejecting the null hypothesis when it is false.

Table 2.6.1 summarizes the relationship between type I and type II errors when testing hypotheses. Furthermore a graph based on the null hypothesis of no difference is presented in Figure 2.6.1 to illustrate the relationship between α and β (or power) for various β 's under H_0 for various alternatives at $\alpha = 5\%$ and 10% . It can be seen that α decreases as β increases or α increases as β decreases. The only way of decreasing both α and β is to increase the sample size. In clinical trials a typical approach is to first choose a significant level α and then select a sample size to achieve a desired test power. In other words, a sample size is chosen to reduce type II error such that β is within an acceptable range at a prespecified significant level of α . From Table 2.6.1 and Figure 2.6.1 it can be seen that α and β depend on the selected null and alternative hypotheses. As indicated earlier, the hypothesis to be questioned is usually chosen as the null hypothesis. The alternative hypothesis is usually of particular interest to the investigators. In practice, the choice of the

**Figure 2.6.1** Relationship between probabilities of type I and type II errors.

null hypothesis and the alternative hypothesis has an impact on the parameter to be tested. Chow and Liu (2000) indicate that the null hypothesis may be selected based on the importance of the type I error. In either case, however, it should be noted that we will never be able to prove that H_0 is true even though the data fail to reject it.

***p*-Values**

In medical literature *p*-values are often used to summarize results of clinical trials in a probabilistic way. For example, in a study of 10 patients with congestive heart failure, Davis et al. (1979) report that at single daily doses of captopril of 25 to 150 mg, the cardiac index rose from 1.75 ± 0.18 to 2.77 ± 0.39 (mean \pm SD) liters per minute per square meter ($p < 0.001$). Powderly et al. (1995) confirmed that fluconazole was effective in preventing esophageal candidiasis (adjusted relative hazard, 5.8; 95% confidence interval 1.7 to 20.0; $p = 0.004$) in patients with advance human immunodeficiency virus (HIV) infection. In a multicenter trial Coniff et al. (1995) indicate that all active treatments (acarbose, tolbutamide, and acarbose plus tolbutamide) were superior ($p < 0.05$) to placebo in reducing postprandial hyperglycemia and H_bA_{1C} levels in noninsulin-dependent diabetes mellitus (NIDDM) patients. In a study evaluating the rate of bacteriologic failure of amoxicillin-clavulanate in the treatment of acute otitis media, Patel et al. (1995) reveal that the bacteriologic failure was higher in nonwhite boys ($p = 0.026$) and in subjects with a history of three or more previous episodes of acute otitis media ($p = 0.008$). These statements indicated that a difference at least as great as the observed would occur in less than 1 in 100 trials if a 1% level of significance were chosen or in less than 1 in 20 trials if a 5% level of significance were selected provided that the null hypothesis of no difference between treatments is true and the assumed statistical model is correct.

In practice, the smaller the *p*-value shows, the stronger the result is. However, the meaning of a *p*-value may not be well understood. The *p*-value is a measure of the chance that the difference at least as great as the observed difference would occur if the null hypothesis is true. Therefore, if the *p*-value is small, then the null hypothesis is unlikely to be true by chance, and the observed difference is unlikely to occur due to chance alone. The *p*-value is usually derived from a statistical test that depends on the size and direction of the effect (a null hypothesis and an alternative hypothesis). To show this, consider testing the following hypotheses at the 5% level of significance:

$$\begin{aligned} H_0: & \text{There is no difference;} \\ \text{vs. } H_a: & \text{There is a difference.} \end{aligned} \tag{2.6.1}$$

The statistical test for the above hypotheses is usually referred to as a *two-sided test*. If the null hypothesis (i.e., H_0) of no difference is rejected at the 5% level of significance, then we conclude there is a significant difference between the drug product and the placebo. In this case we may further evaluate whether the trial size is enough to effectively detect a clinically important difference (i.e., a difference that will lead the investigators to believe the drug is of clinical benefit and hence of effectiveness) when such difference exists. Typically, the FDA requires at least 80% power for detecting such difference. In other words, the FDA requires there be at least 80% chance of correctly detecting such difference when the difference indeed exists.

Figure 2.6.2 displays the sampling distribution of a two-sided test under the null hypothesis in (2.6.1). It can be seen from Figure 2.6.2 that a two-sided test has equal chance to show

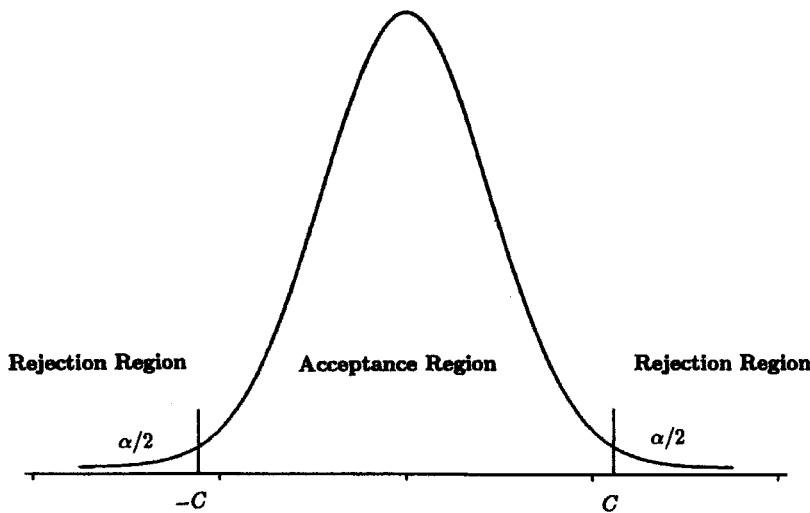


Figure 2.6.2 Sampling distribution of two-sided test.

that the drug is either effective in one side or ineffective in the other side. In Figure 2.6.2, C and $-C$ are critical values. The area under the probability curve between $-C$ and C constitutes the so-called acceptance region for the null hypothesis. In other words, any observed difference in means in this region is a piece of supportive information of the null hypothesis. The area under the probability curve below $-C$ and beyond C is known as the rejection region. An observed difference in means in this region is a doubt of the null hypothesis. Based on this concept, we can statistically evaluate whether the null hypothesis is a true statement. Let μ_D and μ_P be the population means of the primary efficacy variable of the drug product and the placebo, respectively. Under the null hypothesis of no difference (i.e., $\mu_D = \mu_P$), a statistical test, say T can be derived. Suppose that t , the observed difference in means of the drug product and the placebo, is a realization of T . Under the null hypothesis we can expect that the majority of t will fall around the center, $\mu_D - \mu_P = 0$. There is a 2.5% chance that we would see t will fall in each tail. That is, there is a 2.5% chance that t will be either below the critical value $-C$ or beyond the critical value C . If t falls below $-C$, then the drug is worse than the placebo. On the other hand, if t falls beyond C , then the drug is superior to the placebo. In both cases we would suspect the validity of the statement under the null hypothesis. Therefore we would reject the null hypothesis of no difference if

$$t > C \quad \text{or} \quad t < -C.$$

Furthermore we may want to evaluate how strong the evidence is. In this case, we calculate the area under the probability curve beyond the point t . This area is known as the *observed p-value*. Therefore the *p-value* is the probability that a result at least as extreme as that observed would occur by chance if the null hypothesis is true. It can be seen from Figure 2.6.2 that

$$p - \text{value} < 0.05 \quad \text{if and only if} \quad t < -C \quad \text{or} \quad t > C.$$

A smaller *p-value* indicates that t is further away from the center (i.e., $\mu_D - \mu_P = 0$) and consequently provides stronger evidence that supports the alternative hypothesis of

a difference. In practice, we can construct a confidence interval for $\mu_D - \mu_P = 0$. If the constructed confidence interval does not contain 0, then we reject the null hypothesis of no difference at the 5% level of significance. It should be noted that the above evaluations for the null hypothesis reach the same conclusion regarding the rejection of the null hypothesis. However, a typical approach is to present the observed *p*-value. If the observed *p*-value is less than the level of significance, then the investigators would reject the null hypothesis in favor of the alternative hypothesis.

Although *p*-values measure the strength of evidence by indicating the probability that a result at least as extreme as that observed would occur due to random variation alone under the null hypothesis, they do not reflect sample size and the direction of treatment effect. Ware et al. (1992) indicate that *p*-values are a way of reporting the results of statistical analyses. It may be misleading to equate *p*-values with decisions. Therefore, in addition to *p*-values, they recommend that the investigators also report summary statistics, confidence intervals, and the power of the tests used. Furthermore, the effects of selection or multiplicity should also be reported.

Note that when a *p*-value is between 0.05 and 0.01, the result is usually called statistically significant; when it is less than 0.01, the result is often called highly statistically significant.

One-Sided versus Two-Sided Hypotheses

For marketing approval of a drug product, current FDA regulations require that substantial evidence of effectiveness and safety of the drug product be provided. Substantial evidence can be obtained through the conduct of two adequate well-controlled clinical trials. The evidence is considered substantial if the results from the two adequate well-controlled studies are consistent in the positive direction. In other words, both trials show that the drug product is significantly different from the placebo in the positive direction. If the primary objective of a clinical trial is to establish that the test drug under investigation is superior to an active control agent, it is referred to as a superiority trial (ICH E9, 1998). However, the hypotheses given in (2.6.1) do not specify the direction once the null hypothesis is rejected. As an alternative, the following hypotheses are proposed:

$$\begin{aligned} H_0: & \text{There is no difference;} \\ \text{vs. } H_a: & \text{The drug is better than placebo.} \end{aligned} \quad (2.6.2)$$

The statistical test for the above hypotheses is known as *one-sided test*. If the null hypothesis of no difference is rejected at the 5% level of significance, then we conclude that the drug product is better than the placebo and hence is effective. Figure 2.6.3 gives the rejection region of a one-sided test. To further compare a one-sided and a two-sided test, let's consider the level of proof required for marketing approval of a drug product at the 5% level of significance. For a given clinical trial, if a two-sided test is employed, the level of proof required is one out of 40. In other words, at the 5% level of significance, there is 2.5% chance (or one out of 40) that we may reject the null hypothesis of no difference in the positive direction and conclude the drug is effective at one side. On the other hand, if a one-sided test is used, the level of proof required is one out of 20. It turns out that the one-sided test allows more ineffective drugs to be approved because of chance as compared to the two-sided test. As indicated earlier, to demonstrate the effectiveness and safety of a drug product, FDA requires two adequate well-controlled clinical trials be conducted.

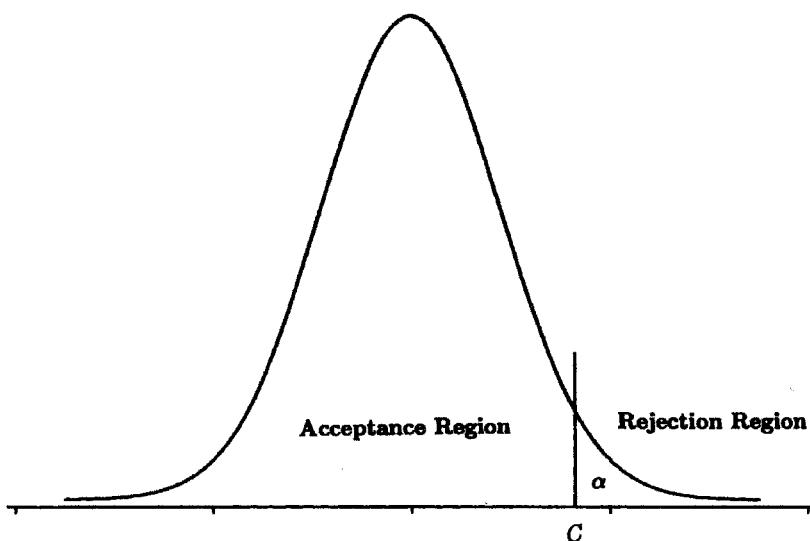


Figure 2.6.3 Sampling distribution of one-sided test.

Then the level of proof required should be squared regardless of which test is used. Table 2.6.2 summarizes the levels of proof required for the marketing approval of a drug product. As Table 2.6.2 indicates, the levels of proof required for one-sided and two-sided tests are one out of 400 and one out of 1600, respectively. Fisher (1991) argues that the level of proof of one out of 400 is a strong proof and is sufficient to be considered as substantial evidence for marketing approval, so the one-sided test is appropriate. However, there is no universal agreement among the regulatory agency (e.g., FDA), academia, and the pharmaceutical industry as to whether a one-sided test or a two-sided test should be used. The concern raised is based on the following two reasons:

1. Investigators would not run a trial if they thought the drug would be worse than the placebo. They would study the drug only if they believe that it might be of benefit.
2. When testing at the 0.05 significance level with 80% power, the sample size required is increased by 27% for the two-sided test as opposed to the one-sided test. As a result there is a substantial impact on cost when a one-sided test is used.

It should be noted that although investigators may believe that a drug is better than the placebo, it is never impossible that such belief might be unexpected (Fleiss, 1987).

Table 2.6.2 Level of Proof Required for Clinical Investigation

Number of Trials	Type of Tests	
	One-Sided	Two-Sided
One trial	1/20	1/400
Two trials	1/400	1/1600

Ellenberg (1990) indicates that the use of a one-sided test is usually a *signal* that the trial has too small a sample size and that the investigators are attempting to squeeze out a significant result by a *statistical maneuver*. These observations certainly argue against the use of one-sided test for the evaluation of effectiveness in clinical trials. Cochran and Cox (1957) suggest that a one-sided test is used when *it is known* that the drug must be at least as good as the placebo, while a two-sided test is used when it is *not known* which treatment is better.

As indicated by Dubey (1991), the FDA tends to oppose the use of a one-sided test. However, this position has been challenged by several drug sponsors on the Drug Efficacy Study Implementation (DESI) drugs at the administrative hearings. As an example, Dubey (1991) points out that several views that favor the use of one-sided test were discussed in an administrative hearing. Some drug sponsors argued that the one-sided test is appropriate in the following situations: (1) where there is truly only concern with outcomes in one tail and (2) where it is completely inconceivable that the results can go in the opposite direction. In this hearing the sponsors inferred that the prophylactic value of the combination drug is greater than that posted by the null hypothesis of equal incidence, and therefore the risk of finding an effect when none in fact exists is located only in the upper tail. As a result a one-sided test is called for. However, the FDA feels that a two-sided test should be applied to account for not only the possibility that the combination drugs are better than the single agent alone at preventing candidiasis but also the possibility that they are worse at doing so.

Dubey's opinion is that one-sided tests may be justified in some situations such as toxicity studies, safety evaluation, analysis of occurrences of adverse drug reactions data, risk evaluation, and laboratory research data. Fisher (1991) argues that one-sided tests are appropriate for drugs that are tested against placebos at the 0.05 level of significance for two well-controlled trials. If, on the other hand, only one clinical trial rather than two is conducted, a one-sided test should be applied at the 0.025 level of significance. However, Fisher agrees that two-sided tests are more appropriate for active control trials.

It is critical to specify hypotheses to be tested in the protocol. A one-sided test or two-sided test can then be justified based on the hypotheses. It should be noted that the FDA is against a post hoc decision to create significance or near significance on any parameters when significance did not previously exist. This critical switch cannot be adequately explained and hence is considered an invalid practice by the FDA. More discussion regarding the use of one-sided test versus two-sided test from the perspectives of the pharmaceutical industry, academe, an FDA Advisory Committee member, and the FDA can be found in Peace (1991), Koch (1991), Fisher (1991), and Dubey (1991), respectively.

2.7 CLINICAL SIGNIFICANCE AND CLINICAL EQUIVALENCE

As indicated in the hypotheses of (2.6.1), the objective of most clinical trials is to detect the existence of predefined clinical difference using a statistical testing procedure such as unpaired two-sample t-test. If this predefined difference is clinically meaningful, then it is of *clinical significance*. If the null hypothesis in (2.6.1) is rejected at the α level of significance, then we conclude that a *statistically significant difference* exists between treatments. In other words, an observed difference that is unlikely to occur by chance alone is considered a statistically significant difference. However, a statistically significant difference depends on the sample size of the trial. A trial with a small sample size usually provides little information regarding the efficacy and safety of the test drug under investigation. On

the other hand, a trial with a large sample size provides substantial evidence of the efficacy of the safety of the test drug product. An observed statistically significant difference, which is of little or no clinical meaning and interpretation, will not be able to address the scientific/clinical questions that a clinical trial was intended to answer in the first place.

The magnitude of a clinically significant difference varies. In practice, no precise definition exists for the clinically significant difference, which depends on the disease, indication, therapeutic area, class of drugs, and primary efficacy and safety endpoints. For example, for antidepressant agents (e.g., Serzone), a change from a baseline of 8 in the Hamilton depression (Ham-D) scale or a 50% reduction from baseline in the Hamilton depression (Ham-D) scale with a baseline score over 20 may be considered of clinical importance. For antimicrobial agents (e.g., Cefil), a 15% reduction in bacteriologic eradication rate could be considered a significant improvement. Similarly, we could also consider a reduction of 10 mm Hg in sitting diastolic blood pressure as clinically significant for ACE inhibitor agents in treating hypertensive patients.

The examples of clinical significance on antidepressant or antihypertensive agents are those of *individual clinical significance*, which can be applied to evaluation of the treatment for individual patients in usual clinical practice. Because individual clinical significance only reflects the clinical change after the therapy, it cannot be employed to compare the clinical change of a therapy to that of no therapy or of a different therapy. Temple (1982) pointed out that in evaluation of one of phase II clinical trials for an ACE inhibitor, although the ACE inhibitor at 150 mg t.i.d. can produce a mean reduction from baseline in diastolic blood pressure of 16 mm Hg, the corresponding mean reduction from baseline for the placebo is also 9 mm Hg. It is easy to see that a sizable proportion of the patients in the placebo group reached the level of individual clinical significance of 10 mm Hg. Therefore, this example illustrates a fact that individual clinical significance alone cannot be used to establish the effectiveness of a new treatment.

For assessment of efficacy/safety of a new treatment modality, it is, within the same trial, compared with either a placebo or another treatment, usually the standard therapy. If the concurrent competitor in the same study is placebo, the effectiveness of the new modality can then be established, based on some primary endpoints, by providing the evidence of an average difference between the new modality and placebo that is larger than some prespecified difference of clinical importance to investigators or to the medical/scientific community. This observed average difference is said to be of the *comparative clinical significance*. The ability of a placebo-controlled clinical trial to provide such observed difference of both comparative clinical significance and statistical significance is referred to as *assay sensitivity*. A similar definition of assay sensitivity is also given in the ICH E10 guidance entitled, *Choice of Control Group in Clinical Trials* (ICH, 1999).

On the other hand, when the concurrent competitor in the trial is the standard treatment or other active treatment, then efficacy of the new treatment can be established by showing that the test treatment is as good as or at least no worse than standard treatment. However, under this situation, the proof of efficacy for the new treatment is based on a crucial assumption that the standard treatment or active competitor has established its own efficacy by demonstrating a difference of comparative clinical significance with respect to placebo in adequate placebo-controlled studies. This assumption is referred to as the *sensitivity-to-drug-effects* (ICH E10, 1999).

Table 2.7.1 presents the results first reported in Leber (1989), which was again used by Temple (1983) and Temple and Ellenberg (2000) to illustrate the issues and difficulties in

Table 2.7.1 Summary of Means of Hamilton Depression Scales of Six Trials Comparing Nomifensine, Imipramine, and Placebo

Study	Common Baseline	Four-week Adjusted Mean (Number of Subjects)		
	Mean	Nomifensine	Imipramine	Placebo
R301	23.9	13.4(33)	12.8(33)	14.8(36)
G305	26.0	13.0(39)	13.4(30)	13.9(36)
C311(1)	28.1	19.4(11)	20.3(11)	18.9(13)
V311(2)	29.6	7.3(7)	9.5(8)	23.5(7)
F313	37.6	21.9(7)	21.9(8)	22.0(8)
K317	26.1	11.2(37)	10.8(32)	10.5(36)

Source: Temple and Ellenberg (2000).

evaluating and interpreting the active controlled trials. All six trials compare nomifensine (a test antidepressant) to imipramine (a standard tricyclic antidepressant) concurrently with placebo. The common baseline means and 4-week adjusted group means based on the Hamilton depression scale are given in Table 2.7.1. Except for trial V311(2), based on the Hamilton depression scale, both nomifensine and imipramine showed more than 50% mean reduction. However, magnitudes of average reduction on the Hamilton depression scale at 4 weeks for the placebo are almost the same as the other two active treatments for all five trials. Therefore, these five trials do not have assay sensitivity. It should be noted that trial V311(2) is the smallest trial, with a total sample size only of 22 patients. However, it was the only trial in Table 2.7.1 that demonstrates that both nomifensine and imipramine are better than placebo in the sense of both comparative clinical significance and statistical significance.

Basically, there are four different outcomes for significant differences in a clinical trial. The result may show that (1) the difference is both statistically and clinically significant, (2) there is a statistically significant difference yet the difference is not clinically significant, (3) the difference is of clinical significance yet not statistically significant, and (4) the difference is neither statistically significant nor clinically significant. If the difference is both clinically and statistically significant or if it is neither clinically nor statistically significant, then there is no confusion. The conclusion can be drawn based on the results from the clinical data. However, in many cases a statistically significant difference does not agree with the clinically significant difference. For example, a statistical test may reveal that there is a statistically significant difference. However, if the difference is too small (it may be due to a unusually small variability or a relatively large sample size) to be of any clinical importance, then it is not clinically significant. In this case a small *p*-value may be instrumental in concluding the effectiveness of the treatment. On the other hand, the result may indicate that there is a clinically significant difference but the sample size is too small (or variability is too large) to claim a statistically significant difference. In this case the evidence of effectiveness is not substantial due to a large *p*-value. This inconsistency has created confusion/arguments among clinicians and biostatisticians in assessment of the efficacy and safety of clinical trials.

As indicated earlier, for the assessment of efficacy and safety of a drug product, a typical approach is to first demonstrate that there is a statistically significant difference between the

drug products in terms of some clinical endpoints by testing hypotheses (2.6.1) repeated below:

$$\begin{aligned} H_0: & \text{ There is no difference;} \\ \text{vs. } H_a: & \text{ There is a difference.} \end{aligned}$$

Equivalently

$$\begin{aligned} H_0: & \mu_D = \mu_P \\ \text{vs. } H_a: & \mu_D \neq \mu_P, \end{aligned}$$

where μ_D and μ_P are the means of the primary clinical endpoint for the drug product and the placebo, respectively. If we reject the null hypothesis of no difference at the α level of significance, then there is a statistically significant difference between the drug product and the placebo in terms of the primary clinical endpoint. We then further evaluate whether there is sufficient power to correctly detect a clinically significant difference. If it does, then we can conclude that the drug product is effective and safe. Note that the above hypotheses are known as *point hypotheses*. In practice, it is recognized that no two treatments will have exactly the same mean responses. Therefore, if the mean responses of the two treatments differ by less than a meaningful limit (i.e., a clinically important difference), the two treatments can be considered clinically equivalent. Based on this idea, Schuirmann (1987) first introduces the use of *interval hypotheses* for assessing bioequivalence. The interval hypotheses for clinical equivalence can be formulated as

$$\begin{aligned} H_0: & \text{ The two drugs are not equivalent;} \\ \text{vs. } H_a: & \text{ The two drugs are equivalent.} \end{aligned} \tag{2.7.1}$$

Or put differently,

$$\begin{aligned} H_0: & \mu_A - \mu_B \leq L \quad \text{or} \quad \mu_A - \mu_B \geq U; \\ \text{vs. } H_a: & L < \mu_A - \mu_B < U, \end{aligned}$$

where μ_A and μ_B are the means of the primary clinical endpoint for drugs A and B, respectively, and L and U are some clinically meaningful limits. The concept and interval hypotheses (2.7.1) is to show equivalence by rejecting the null hypothesis of inequivalence. The above hypotheses can be decomposed into two sets of one-sided hypotheses

$$\begin{aligned} H_{01}: & \text{ Drug } A \text{ is superior to drug } B \text{ (i.e., } \mu_A - \mu_B \geq U\text{);} \\ \text{vs. } H_{a1}: & \text{ Drug } A \text{ is not superior to drug } B; \end{aligned}$$

and

$$\begin{aligned} H_{02}: & \text{ Drug } A \text{ is inferior to drug } B \text{ (i.e., } \mu_A - \mu_B \leq L\text{);} \\ \text{vs. } H_{a2}: & \text{ Drug } A \text{ is not inferior to drug } B \end{aligned}$$

The first set of hypotheses is to verify that drug A is not superior to drug B , while the second set of hypotheses is to verify that drug A is not worse than drug B . A relatively large or

small observed difference may refer to the concern of the comparability between the two drug products. Therefore the rejection of H_{01} and H_{02} will lead to the conclusion of clinical equivalence. This is equivalent to rejecting H_0 in (2.7.1). In practice, if L is chosen to be $-U$, then we can conclude clinical equivalent if

$$|\mu_A - \mu_B| < \Delta,$$

where $\Delta = U = -L$ is the clinically significant difference. For example, for the assessment of bioequivalence between a generic drug product and an innovator drug product (or reference drug product), the bioequivalence limit Δ is often chosen to be 20% of the bioavailability of the reference product. In other words, in terms of the ratio of means μ_A/μ_B , the limits become $L = 80\%$ and $U = 120\%$. When log-transformed data are analyzed, the FDA suggests using $L = 80\%$ and $U = 125\%$. More detail on the assessment of bioequivalence between drug products can be found in Chow and Liu (2000).

When two drugs are shown to be clinically equivalent, they are comparable to each other. Consequently they can be used as substitutes for each other. It should be noted that there is difference between the assessment of a possible difference and equivalence. Hypotheses (2.6.1) are set for assessment of a possible difference between treatments, while hypotheses (2.7.1) are for the assessment of equivalence. The demonstration of equality does not necessarily imply equivalence. This is because the selected sample size for testing equality may not be sufficient for assessing the equivalence. Besides, when we fail to reject the null hypothesis of equality, it does not imply that the two treatments are equivalent, even if there is sufficient power for the detection of a clinically significant difference.

Note that the current FDA regulations do not allow the sponsors to establish clinical equivalence/noninferiority based on clinical trials designed for the detection of existence of treatment differences. On the other hand, the ICH E9 guideline also stressed that it is inappropriate to conclude equivalence/noninferiority based on observing a statistically nonsignificant test result for null hypothesis (2.6.1) that there is no difference between the investigational drug and the active competitor. Clinical equivalence/noninferiority between two drug products must be established based on the interval hypothesis, as described in (2.7.1). Confidence approach in general is used to establish the clinical equivalence/noninferiority. If the entire confidence interval for the average difference between the investigator product and active competitor is within some prespecified equivalence limit, clinical equivalence is inferred. Clinical noninferiority is concluded if an upper one-sided confidence limit is smaller than the prespecific limit. From the above discussion, the sample size determination for equivalence/noninferiority trials should specify the value of Δ , which is the largest difference between the investigational product and the active competitor that can be judged as clinically acceptable. In addition, the power for concluding equivalence/noninferiority using a prespecified value of Δ should be given. For testing interval hypothesis, several statistical procedures have been proposed. See, for example, Blackwelder (1982), Wellek (1993), Jennison and Turnbull (1993), and Liu (1995a). Equivalence/noninferiority trials without inclusion of a placebo group are not internally valid and rely on external validation of the assumed sensitivity-to-drug effects. Furthermore, selection of equivalence limits is also a very controversial issue and recently sparks heated arguments between the sponsors and the regulatory agencies. See Jones et al. (1996), Rohmel (1998), Ebbutt and Firth (1998), Fisher et al. (2001), Fleming (2000), Siegel (2000), Temple and Ellenberg (2000), and Ellenberg and Temple (2000). More details are given in Chapter 7.

2.8 REPRODUCIBILITY AND GENERALIZABILITY

As indicated in the previous chapter, for marketing approval of a new drug product, the U.S. FDA requires that substantial evidence of the effectiveness and safety of the drug product be provided through the conduct of at least two adequate and well-controlled clinical trials. The purpose of requiring at least two pivotal clinical trials is not only to assure the reproducibility, but also to provide valuable information regarding generalizability. Chow and Shao (2002) define reproducibility as to (1) whether the clinical results in the same target patient population are reproducible from one location (e.g., study site) to another within the same region (e.g., the United States of America, European Union, or Asian Pacific region) or (2) whether the clinical results are reproducible from one region to another region in the same target patient population. Generalizability is referred to as (1) whether the clinical results can be generalized from the target patient population (e.g., adult) to another similar but slightly different patient population (e.g., elderly) within the same region or (2) whether the clinical results can be generalized from the target patient population (e.g., white) in one region to a similar but slightly different patient population (e.g., Asian) in another region. In what follows, we will provide the concept of reproducibility and generalizability for providing substantial evidence in clinical research and development.

Reproducibility

In clinical research, two questions are commonly asked. First, what is the chance that we will observe a negative result in a future clinical study under the same study protocol given that positive results have been observed in the two pivotal trials? In practice, two positive results observed from the two pivotal trials, which have fulfilled the regulatory requirement for providing substantial evidence, may not guarantee that the clinical results are reproducible in a future clinical trial with the same study protocol with a high probability. This is very likely, especially when the positive results observed from the two pivotal trials are marginal (i.e., their p -values are close to but less than the level of significance). Second, it is often of interest to determine whether a large clinical trial that produced positive clinical results can be used to replace two pivotal trials for providing substantial evidence for regulatory approval. Although the U.S. FDA requires at least two pivotal trials be conducted for providing substantial evidence regarding the effectiveness and safety of the drug product under investigation for regulatory review, under the circumstances, the FDA Modernization Act (FDAMA) of 1997 includes a provision (Section 115 of FDAMA) to allow data from one adequate and well-controlled clinical trial investigation and confirmatory evidence to establish effectiveness for risk/benefit assessment of drug and biological candidates for approval. To address the above two questions, Shao and Chow (2002) suggested evaluating the probability of observing a positive result in a future clinical study with the same study protocol, given that a positive clinical result has been observed.

Let H_0 and H_a be the null hypothesis and the alternative hypothesis of (2.6.1). Thus, the null hypothesis is that there is no difference in mean response between a test drug and a control (e.g., placebo). Suppose that the null hypothesis is rejected if and only if $|T| > C$, where C is a positive known constant and T is a test statistic, which is usually related to a two-sided alternative hypothesis. In statistical theory, the probability of observing a significant clinical result when H_a is indeed true is referred to as the power of the test procedure. If the statistical model under H_a is a parametric model, then the power can be evaluated at θ , where θ is an unknown parameter or vector of parameters. Suppose now

that one clinical trial has been conducted and the result is significant. Then, what is the probability that the second trial will produce a significant result, i.e., the significant result from the first trial is reproducible? Statistically, if the two trials are independent, the probability of observing a significant result in the second trial when H_a is true is the same as that of the first trial regardless of whether the result from the first trial is significant. However, it is suggested that information from the first clinical trial should be used in the evaluation of the probability of observing a significant result in the second trial. This leads to the concept of reproducibility probability (Shao and Chow, 2002).

In general, the reproducibility probability is a person's subjective probability of observing a significant clinical result from a future trial, when he/she observes significant results from one or several previous trials. Goodman (1992) considered the reproducibility probability as the power of the trial (evaluated at θ) by simply replacing θ with its estimate based on the data from previous trials. In other words, the reproducibility probability can be defined as an *estimated power* of the future trial using the data from previous studies. Shao and Chow (2002) studied how to evaluate the reproducibility probability using this approach under several study designs for comparing means with both equal and unequal variances. When the reproducibility probability is used to provide substantial evidence of the effectiveness of a drug product, the estimated power approach may produce an optimistic result. Alternatively, Shao and Chow (2002) suggested that the reproducibility probability be defined as a lower confidence bound of the power of the second trial. In addition, they also suggested a more sensible definition of reproducibility probability using the Bayesian approach. Under the Bayesian approach, the unknown parameter θ is a random vector with a prior distribution, say, $\pi(\theta)$, which is assumed known. Thus, the reproducibility probability can be defined as the conditional probability of $|T| > C$ in the future trial, given the data set.

In practice, the reproducibility probability is useful when the clinical trials are conducted sequentially. It provides important information for regulatory agencies in determining whether it is necessary to require the second clinical trial when the result from the first clinical trial is strongly significant. To illustrate the concept of reproducibility probability, reproducibility probabilities for various values of $|T(x)|$ with $n = 30$ are given in Table 2.8.1. Table 2.8.1 suggests that it is not necessary to conduct the second trial if the observed p -value of the first trial is less than or equal to 0.001 because the reproducibility probability is about 0.91. On the other hand, even when the observed p -value is less than the 5% level of significance, say, the observed p -value is less than or equal to

Table 2.8.1 Reproducibility Probability \hat{P}

$ T(x) $	Known σ^2		Unknown σ^2 ($n = 30$)	
	p -value	\hat{P}	p -value	\hat{P}
1.96	0.050	0.500	0.060	0.473
2.05	0.040	0.536	0.050	0.508
2.17	0.030	0.583	0.039	0.554
2.33	0.020	0.644	0.027	0.614
2.58	0.010	0.732	0.015	0.702
2.81	0.005	0.802	0.009	0.774
3.30	0.001	0.910	0.003	0.890

Source: Chow and Shao (2002).

0.01, a second trial is recommended because the reproducibility probability may not reach the level of confidence for the regulatory agency to support the substantial evidence of effectiveness of the drug product under investigation. When the second trial is necessary, the reproducibility probability can be used for sample size adjustment of the second trial. More details regarding sample size calculation based on reproducibility can be found in Shao and Chow (2002) and Chow et al. (2003) and are discussed in Section 7.9.

Generalizability

As discussed above, the concept of reproducibility is to evaluate whether clinical results observed from the *same* targeted patient population are reproducible from study site to study site within the same region or from region to region. In clinical development, after the drug product has been shown to be effective and safe with respect to the targeted patient population, it is often of interest to determine how likely the clinical results can be reproducible to a *different but similar* patient population with the same disease. We will refer to the reproducibility of clinical results in a different but similar patient population as the generalizability of the clinical results. For example, if the approved drug product is intended for the adult patient population, it is often of interest to study the effect of the drug product on a different but similar patient population, such as the elderly or pediatric patient population with the same disease. In addition, it is also of interest to determine whether the clinical results can be generalized to patient populations with ethnic differences. Similarly, Shao and Chow (2002) proposed to consider the so-called generalizability probability, which is the reproducibility probability with the population of a future trial slightly deviated from the targeted patient population of previous trials, to determine whether the clinical results can be generalized from the targeted patient population to a different but similar patient population with the same disease.

In practice, the response of a patient to a drug product under investigation is expected to vary from patient to patient, especially from patients from the target patient population to patients from a different but similar patient population. The responses of patients from a different but similar patient population could be different from those from the target patient population. As an example, consider a clinical trial, which was conducted to compare the efficacy and safety of a test drug with an active control agent for treatment of schizophrenia patients and patients with schizoaffective disorder. The primary study endpoint is the positive and negative symptom score (PANSS). The treatment duration of the clinical trial was 1 year with a 6-month follow-up. Table 2.8.2 provides summary statistics of PANSS by race. As it can be seen from Table 2.8.2, the means and standard deviations of PANSS are different across different races. Oriental patients tend to have higher PANSS with less variability as compared to those in white patients. Black patients seem to have lower PANSS with less variability at both baseline and endpoint. Thus, it is of interest to determine that the observed clinical results can be generalized to a different but similar patient population such as black or Oriental.

Chow (2001) indicated that the responses of patients from a different but similar patient population could be described by the changes in mean and variance of the responses of patients from the target patient population. Consider a parallel-group clinical trial comparing two treatments with population means μ_1 and μ_2 and an equal variance σ^2 . Suppose that in the future trial, the population mean difference is changed to $\mu_1 - \mu_2 + \varepsilon$ and the population variance is changed to $C^2\sigma^2$, where $C > 0$. The signal-to-noise ratio for the population difference in the previous trial is $|\mu_1 - \mu_2|/\sigma$, whereas the signal-to-noise ratio

Table 2.8.2 Summary Statistics of PANSS

Race		Baseline			Endpoint		
		All Subjects	Test	Active Control	All Subjects	Test	Active Control
All Subjects	N	364	177	187	359	172	187
	Mean	66.3	65.1	67.5	65.6	61.8	69.1
	S.D.	16.85	16.05	17.54	20.41	19.28	20.83
	Median	65.0	63.0	66.0	64.0	59.0	67.0
	Range	(30-131)	(30-115)	(33-131)	(31-146)	(31-145)	(33-146)
White	N	174	81	93	169	77	92
	Mean	68.6	67.6	69.5	69.0	64.6	72.7
	S.D.	17.98	17.88	18.11	21.31	21.40	20.64
	Median	65.5	64.0	66.0	66.0	61.0	70.5
	Range	(30-131)	(30-115)	(33-131)	(31-146)	(31-145)	(39-146)
Black	N	129	67	62	129	66	63
	Mean	63.8	63.3	64.4	61.7	58.3	65.2
	S.D.	13.97	12.83	15.19	18.43	16.64	19.64
	Median	64.0	63.0	65.5	61.0	56.5	66.0
	Range	(34-109)	(38-95)	(34-109)	(31-129)	(31-98)	(33-129)
Oriental	N	5	2	3	5	2	3
	Mean	71.8	72.5	71.3	73.2	91.5	61.0
	S.D.	4.38	4.95	5.03	24.57	20.51	20.95
	Median	72.0	72.5	72.0	77.0	91.5	66.0
	Range	(66-76)	(69-76)	(66-76)	(38-106)	(77-106)	(38-79)
Hispanic	N	51	24	27	51	24	27
	Mean	64.5	61.4	67.3	64.6	61.9	67.1
	S.D.	18.71	16.78	20.17	20.60	16.71	23.58
	Median	63.0	60.0	68.0	66.0	59.5	67.0
	Range	(33-104)	(35-102)	(33-104)	(33-121)	(33-90)	(33-121)

Table 2.8.3 Effects of Changes in Mean and Standard Deviation (ε and C) on Δ

$ \varepsilon/(\mu_1 - \mu_2) $	C	Range of Δ
<5%	0.8	1.188–1.313
	0.9	1.056–1.167
	1.0	0.950–1.050
	1.1	0.864–0.955
	1.2	0.731–0.808
	1.3	0.731–0.808
	1.4	0.679–0.750
	1.5	0.633–0.700
$\geq 5\%$ but <10%	0.8	1.125–1.375
	0.9	1.000–1.222
	1.0	0.900–1.100
	1.1	0.818–1.000
	1.2	0.750–0.917
	1.3	0.692–0.846
	1.4	0.643–0.786
	1.5	0.600–0.733
$\geq 10\%$ but <20%	0.8	1.000–1.500
	0.9	0.889–1.333
	1.0	0.800–1.200
	1.1	0.727–1.091
	1.2	0.667–1.000
	1.3	0.615–0.923
	1.4	0.571–0.857
	1.5	0.533–0.800

Source: Shao and Chow (2002).

for the population difference in the future trial is

$$\frac{|\mu_1 - \mu_2 + \varepsilon|}{C\sigma} = \frac{|\Delta(\mu_1 - \mu_2)|}{\sigma}$$

where

$$\Delta = \frac{1 + \varepsilon/(\mu_1 - \mu_2)}{C}$$

is a measure of change in the signal-to-noise ratio for the population difference. Note that the above can be expressed by $|\Delta|$ multiplying the *effect size* of the first trial. As a result, Chow et al., (2002) refer to Δ as sensitivity index, which is useful when assessing similarity in bridging studies. For most practical problems, $|\varepsilon| < |\mu_1 - \mu_2|$ and thus $\Delta > 0$. Table 2.8.2 gives an example on the effects of changes of ε and C and Δ .

If the power for the previous trial is $p(\theta)$, then the power for the future trial is $p(\Delta\theta)$. Suppose that Δ is known. As discussed earlier, the generalizability probability is given by \hat{P}_Δ , which can be obtained by simply replacing $T(x)$ with $\Delta T(x)$. Under the Bayesian approach, the generalizability probability can be obtained by replacing $p(\delta/u)$ with $p(\Delta\delta/u)$.

In practice, the generalizability probability is useful when assessing similarity between clinical trials conducted in different regions (e.g., Europe and the United States of America or the United States of America and Asian Pacific region). It provides important information for local regulatory health authorities in determining whether it is necessary to require a bridging clinical study based on the analysis of the sensitivity index for assessment of possible difference in ethnic factors (Chow et al., 2002). When a bridging study is deemed necessary, the assessment of generalizability probability based on the sensitivity index can be used for sample size adjustment of the bridging clinical study. More details are provided in Section 7.7 (see also Chow et al., 2003b).

3

BASIC DESIGN CONSIDERATIONS

3.1 INTRODUCTION

In the clinical development of a drug product, clinical trials are often conducted to address scientific and/or medical questions regarding the drug product in treatment of a specific patient population with certain diseases. At the planning stage of a clinical trial, it is therefore important to define “**What is the question?**” Defining “What is the question?” helps determine the study objective(s) and consequently helps set up appropriate hypotheses for scientific evaluation. The next question is “**How to answer the question?**” It is important that the intended clinical trial provide an unbiased and valid scientific evaluation of the question. Temple (1982) indicates that two kinds of difficulties are often encountered when the clinical scientists attempt to identify/answer pertinent scientific questions by conducting well-controlled clinical trials. The first difficulty is that individual studies may be designed without careful attention to the questions they really are capable of answering. Consequently the trial is either a useless trial that answers no question at all or it is a trial that answers some other question (but not the one intended) or only part of the intended question. Second, the total package of studies may be designed without a thoughtful consideration of all the questions that are pertinent. There are practical limitations on the number of studies that can reasonably be expected; nevertheless, it seems possible that more of the pertinent questions can be answered without any increase in the total number of patients exposed in clinical trials.

To best answer scientific and/or medical questions through clinical trials, the FDA suggests that an overall study plan and design be briefly but clearly described in the protocol of the intended clinical trial. **A thoughtful and well-organized protocol includes study objective(s), study design, patient selection criteria, dosing schedules, statistical methods, and**

other medical related details. As a result, “How to choose an appropriate study design?” and “How to analyze the collected clinical data using valid statistical methods?” have become two important aspects of a clinical trial plan. These two aspects are closely related to each other since statistical methods for data analysis depend on the design employed. Generally speaking, meaningful conclusions can only be drawn based on data collected from a valid scientific design using appropriate statistical methods. Therefore, the selection of an appropriate study design is important in order to provide an unbiased and scientific evaluation of the scientific and/or medical questions regarding the study drug. Before a study design is chosen, some basic design considerations such as goals of clinical trials, patient selection, randomization and blinding, the selection of control(s), and some statistical issues must be considered to justify the use of statistical analyses. In this chapter our efforts will be directed to the objectives of clinical trials and the selection of patients for clinical trials, the selection of control(s), statistical considerations, and some other related issues. Randomization and blinding will be discussed in detail in the next chapter. Several commonly employed designs in clinical trials are reviewed in Chapter 5.

In the next section, the goals of clinical research and the manners to specify the objectives of a clinical trial are addressed. Issues in defining the target patient population and selecting patients for a clinical trial are discussed in Section 3.3. In Section 3.4, we discuss the selection of control(s) in clinical trials. Some statistical considerations regarding clinical evaluation of efficacy and safety, sample size estimation, interim analysis and data monitoring, and statistical and clinical inference are given in Section 3.5. Section 3.6 contains some specific issues related to designing a clinical trial such as single site versus multi-sites, treatment duration, patient compliance, and missing value and dropout. A brief concluding discussion is given in the last section.

3.2 GOALS OF CLINICAL TRIALS

The ultimate goal of clinical research is to obtain an unbiased inference with possibly best precision in order to scientifically address the clinical questions regarding the study drug under investigation with respect to a target patient population. As indicated by Lachin (2000), the meaning of an unbiased trial is two-fold. First, the estimated treatment effect between the investigational drug and a control is unbiased. Second, the statistical testing procedure for detecting a treatment effect is also unbiased in the sense that the false-positive rate (i.e., type I error rate) for concluding the existence of a treatment effect is controlled at a prespecified nominal level of significance. On the other hand, the best precision of an inference implies that the variability of the estimated treatment effect based on the data obtained from a clinical trial is the smallest. Consequently, it has the highest likelihood to reproduce its results in the same target patient population and to generalize its results to a different patient population. All of the methodologies introduced and illustrated in this book are to minimize bias (or increase accuracy) and to maximize precision (reliability) associated with a clinical trial.

In clinical research, however, how to develop/formulate a feasible and yet scientifically valid set of important clinical/medical questions to be addressed by the intended clinical trial is probably one of the most difficulties commonly encountered for achieving the goals of minimizing bias and variability. Once the clinical/medical questions have been clearly stated, necessary resources such as the number of subjects, study duration, study endpoints for evaluation of the study drug, facility/equipment, and clinical personnel can be determined in

order to provide an accurate and reliable statistical/clinical inference for addressing these questions. The most commonly seen mistake in the conduct of clinical trials is that the investigator(s) often attempts to answer all possible questions with respect to a certain therapeutic area in a single trial regardless the size of the trial. As a result, the objectives of the study may be too ambitious and/or too unspecific to be answered by the limited clinical data observed from the trial at the end of the trial. In addition, the study may require too much resource and/or too long to complete, which might be beyond the capacity of the sponsor and/or funding agencies of the trial. Hence, we define the objective of a clinical trial as a statement regarding a set of clinical/medical questions that are clear, concise, precise, scientifically valid, and quantitative that can be easily translated into hypotheses.

For illustration purposes, in what follows, three examples regarding the objectives of clinical trials that are commonly seen in clinical/medical literature are provided. The first two examples are for drug evaluation, and the third example is related to a smoking prevention trial.

Example 3.2.1

The objective of this study is to evaluate the efficacy and safety of the test drug under investigation with a placebo in the treatment of postmenopausal women with osteoporosis.

Example 3.2.2

The objective of the trial is to compare the efficacy and safety of the test drug under investigation and an active control agent given on demand with a placebo in males with erectile dysfunction.

Example 3.2.3

The objective of this smoking prevention trial is to address the scientific question: To what extent can the school-based tobacco usage prevention intervention deter tobacco usage, by both genders, throughout and beyond high school?

Note that the objectives given in the above examples do not provide the study endpoints for evaluation of the efficacy and safety of the test drug under investigation. Example 3.2.2 is a three-arm active control trial with a placebo group. With respect to efficacy, it should have at least two objectives: (1) confirmation of assay validity by demonstrating a superior efficacy of the active competitor over the placebo, and (2) providing evidence on the superiority or equivalence/noninferiority of the test drug under investigation with the active competitor. If there are multiple objectives for a clinical trial, one should prioritize these objectives as the primary objective(s) and secondary objectives. Example 3.2.1 can be rewritten in a more specific manner, as demonstrated in the following example.

Example 3.2.4

Primary Objectives

1. This trial is a randomized, double-blind, placebo-controlled trial conducted in x centers to evaluate the efficacy based on bone mineral density (BMD) of the test drug under

investigation at dose y, frequency z, compared to a placebo, in the treatment of postmenopausal women with osteoporosis.

- 2.** This trial is a randomized, double-blind, placebo-controlled trial conducted in x centers to evaluate the safety of the test drug under investigation at dose y, frequency z, compared to a placebo, in the treatment of postmenopausal women with osteoporosis.

Secondary Objectives

- 1.** To evaluate the effectiveness of the test drug under investigation on the incidence of vertebral fractures.
- 2.** To evaluate the effectiveness of the test drug under investigation on biochemical markers of bone turnover.

As illustrated in Example 3.2.4, the efficacy of the test drug is evaluated by the primary efficacy endpoint of BMD and two secondary efficacy endpoints of the incidence of vertebral fractures and biochemical marker for assessment of bone loss. However, the direction of hypothesis that the investigator would like to verify or confirm in the trial is not addressed in the example. In addition, safety parameters are not specified in the objectives either. Example 3.2.5 provides the objectives for evaluation of an investigational drug in treatment of patients with chronic obstructive pulmonary disease (COPD).

Example 3.2.5

Primary Objectives

This trial is a randomized, double-blind, placebo-controlled trial conducted in x centers to assess the efficacy of the test drug under investigation at dose y, frequency z for a 52-week treatment of patients with COPD in terms of

- 1.** The reduction in the risk of exacerbation of COPD for patients receiving the investigational drug as compared to the placebo.
- 2.** A superior pulmonary function as assessed by the forced expiratory volume in one second (FEV_1) for patients receiving the investigational drug as compared to the placebo.

Secondary Objectives

The secondary objectives are to assess the safety and tolerability of the investigational drug at dose y, frequency z versus placebo by the evaluation of the incidence rates of adverse events and laboratory parameters over a 52-week treatment of patients with COPD.

The primary objectives in Example 3.2.5 clearly state that the purpose of the trial is to show a superior efficacy of the investigational drug over the placebo. Furthermore, a superior efficacy for the investigational drug is established if the superiority in both the risk of the exacerbation of COPD and the pulmonary function can be proven. Therefore, the superiority of the investigational drug is based on both primary study endpoints. In addition, secondary objectives also state that safety and tolerability will be evaluated based on adverse events and laboratory parameters.

Example 3.2.6**Primary Objectives**

This is a randomized, parallel-group trial to demonstrate that the one-year survival of the patients with pretreated advanced (Stage IIIB/IV) non-small-cell lung cancer (NSCLC) receiving the oral investigational drug is not inferior to those receiving intravenous (IV) docetaxel.

Secondary Objectives

Secondary objectives of the trial are to evaluate overall survival, time to progression, response rate, time to response, improvement in quality of life, and qualitative and quantitative toxicities.

Example 3.2.6 provides a typical example of the objectives of a noninferiority trial. In a noninferiority trial, the purpose of the trial is to show that the efficacy of the investigational drug is no worse than (or at least as effective as) the active competitor. In addition to the efficacy, it is often of interest to show that the test drug has a better safety profile in a noninferiority trial. Therefore, it is suggested that the objective of showing a superior safety should be clearly stated as one of the primary objectives. In practice, it is then preferred that both objectives regarding the efficacy and safety must be achieved by providing substantial evidence for regulatory review. In general, the study endpoints for addressing the primary and secondary objectives are different but may be correlated. In this case, sometimes, composite endpoints are used to evaluate the efficacy of the test drug under investigation. For the primary and secondary objectives, these composite endpoints may consist of different combinations of the occurrence of different events. In some cases, a clinical trial may be intended to explore the efficacy and safety of the test drug in certain predefined subgroups. These objectives can be stated as tertiary objectives.

Example 3.2.7**Primary Objectives**

This is a randomized, parallel-group study for determining whether the efficacy of the test drug is not inferior to that of the active competitor for the prevention of all stroke (fatal and nonfatal) and systemic embolic events in patients with arterial fibrillation (AF).

Secondary Objectives

Secondary objectives include the following:

1. To compare the efficacy of the test drug to that of the active competitor in terms of the composite endpoint of the prevention of death, nonfatal stroke, nonfatal systemic embolic events, and nonfatal myocardial infarction in patients with arterial fibrillation (AF).
2. To compare the safety of the test drug to that of the active competitor with major and major bleeding events and treatment discontinuation as the primary safety endpoints.

Tertiary Objectives

The following tertiary objectives are of interest:

1. To compare the efficacy of the test drug to that of the active competitor for the prevention of all strokes with a poor outcome (defined as a Modified Rankin score of 3 or more at three-months post-stroke or a Barthel score less than 60 at three-months post-stroke).
2. To compare the efficacy of the test drug to that of the active competitor for the prevention of all strokes and systemic embolic events in patients of 75 years of age or above with AF and to compare this with patients below the age of 75 years.

The efficacy endpoints in Example 3.2.7 are the occurrence of prespecified events. For the primary objectives, these events include all strokes and systemic embolic events while the secondary efficacy endpoints consist of death, nonfatal stroke, nonfatal systemic embolic events, and nonfatal myocardial infarction. As a result, the primary and secondary efficacy endpoints contain some overlapping events such as nonfatal strokes. The first tertiary objective is to evaluate the performance of the test drug on the prespecified subgroup based on the primary efficacy endpoint for the patients with a poor outcome. On the other hand, the second tertiary objective is to detect a possible interaction between the age (equal to 75 years or older vs. younger than 75 year) and treatment based on the primary efficacy endpoint.

3.3 TARGET POPULATION AND PATIENT SELECTION

As was indicated earlier, one of the primary objectives of a clinical trial is to provide an accurate and reliable clinical evaluation of a study drug for a target patient population with certain diseases. In practice, statistical and clinical inference are usually drawn based on a representative sample (a group of patients to be enrolled in the trial) selected from the target patient population of the clinical trial. A representative sample provides the clinician with the ability to generalize the findings of the study. Therefore, selecting patients for a clinical trial plays an important role to best answer the scientific and/or medical questions of interest regarding the study drug. Basically selecting patients for a clinical trial involves two steps. First, we need to define the target patient population. Patients are then selected from the target patient population for the clinical trial. For a given disease, the target patient population is often rather heterogeneous with respect to patient characteristics and the severity of the disease. The heterogeneity of the target patient population can certainly decrease the accuracy, reliability, and the generalization of the findings of the study. In clinical trials, the target patient population usually involves various sources of expected and unexpected biases and variabilities. For example, bias and variability due to differences in patient demographic characteristics such as age, sex, height, weight, and functional status are expected. Bias and variability caused by changes in disease status and concomitant therapies are unexpected. These sources of biases and variabilities will not only decrease the accuracy and reliability of the observed clinical results but also limit the clinician's ability to generalize the findings of the study. For good clinical practice, it is therefore desirable to define the target patient population in such a way that it is a homogeneous as possible with respect to these patient characteristics in order to reduce bias and to minimize variability. For this purpose, Section 314.166 of CFR also requires that the

method for selection of patients in clinical trials provide adequate assurance that the selected patients have the disease and condition being studied.

For patient selection, Weintraub and Calimlim (1994) classify patients into two categories. These two categories are inpatients for short-term hospital studies and outpatients for chronic conditions. Different concerns/considerations may be raised depending on which type of patients are intended for the clinical trials. In this section, we will focus on a general concept for selecting patients for a clinical trial which includes the development of eligibility criteria, selection process, and ethical considerations.

Eligibility Criteria

In clinical trials a set of eligibility criteria is usually developed to define the target patient population from which qualified (or eligible) patients can be recruited to enroll the studies. Typically a set of eligibility criteria consists of a set of inclusion criteria and a set of exclusion criteria. The set of inclusion criteria is used to roughly outline the target patient population, while the set of exclusion criteria is used to fine-tune the target patient population by removing the expected sources of variabilities. To be eligible for the intended study, patients must meet *all* the inclusion criteria. Patients meeting *any* of the exclusion criteria will be excluded from the study. Eligibility criteria should be developed based on patient characteristics, diagnostic criteria, treatment duration, and the severity of the disease.

Before a set of well-defined eligibility criteria can be developed, it is necessary to have a clear understanding of the study medicine and the indication it is intended for. For example, some medicines are intended for specific patient population (e.g., female, children, or elderly) with a certain disease. The inclusion criteria usually describe the target patient population based on the diagnosed symptoms or history of the intended disease. Patients who have history of hypersensitivity to the study medicine, treatment-resistance, disease changes, and/or concurrent diseases requiring treatments are usually excluded from the study. Different eligibility criteria will result in different study patient populations. These differences decrease the ability to apply the study results to any other patient population. In what follows, we provide three examples for the development of eligibility criteria for clinical trials from three major therapeutic areas: anti-infectives, cardiovascular, and central nervous system.

For the first example, consider a clinical trial comparing the clinical and microbiologic efficacy and safety of an antibiotic agent in the treatment of febrile episodes in neutropenic cancer patients. As indicated in the *Guidelines for the Use of Antimicrobial Agents in Neutropenic Patients with Unexplained Fever* (IDSA, 1990), anti-infective drugs have become a standard of medical practice whenever a neutropenic patient becomes febrile. For example, ceftazidime which is a marketed third-generation cephalosporin is indicated for the treatment of febrile episodes in neutropenic cancer patients caused by *Streptococcus spp.*, *Escherichia coli*, *Klebsiella spp.*, *Pseudomonas aeruginosa*, and *Proteus mirabilis* (PDR, 1992). With the more prompt and routine initiation of anti-infective therapy, the microbiologic confirmation of infection has declined such that as many as 50% to 70% of the febrile neutropenic episodes do not have a defined microbial etiology. These patients are categorized as having unexplained fever in which the infection may have been masked by the early introduction of antimicrobial therapy. Since unexplained fever constitutes the majority of febrile neutropenic events, the evaluation of empiric therapy has become more difficult. Consequently, the question raised is how therapy can be adequately assessed when fever is the only evaluable parameter (IHS, 1990). For this reason, the primary clinical endpoint being evaluated in this study is fever. It is suggested an oral temperature greater

Table 3.3.1 Eligibility Criteria for Anti-Infectives Agents

<i>A. Inclusion Criteria</i>
1. Hospitalized patients aged 18 years or older.
2. An oral temperature greater than 38.5°C once or greater than 38°C on two or more occasions during a 12-hour period.
3. Fewer than 500 absolute neutrophils (polymorphonuclear and segmented) per mm ³ , or patients presenting with between 500 and 1000 absolute neutrophils per mm ³ , whose counts are anticipated to fall below 500 per mm ³ within 48 hours because of antecedent therapy.
<i>B. Exclusion Criteria</i>
1. History of hypersensitivity to a cephalosporin or penicillin.
2. Pregnant or breast-feeding.
3. Requiring other systemic antibacterial drugs concomitantly except for intravenous vancomycin.
4. Creatinine clearance ≤ 15 mL/min or requiring hemodialysis or peritoneal dialysis.
5. History of positive antibody test for HIV.
6. A severe underlying disease such as meningitis, osteomyelitis, or endocarditis.
7. Patients undergoing bone marrow transplantation or stem cell harvesting and infusion.
8. Any other condition that in the opinion of the investigator(s) would make the patient unsuitable for enrollment.

than 38.5°C once or greater than 38°C on two or more occasions during a 12-hour period be considered as an inclusion criterion for the study. Note that it may be a concern that the weak antistreptococcal activity of ceftazidime in patients whose infections are frequently caused by gram-positive bacteria. As a result many practitioners have routinely added vancomycin to ceftazidime as initial coverage for the febrile neutropenic patient. Therefore, no other antibacterial agents except intravenous vancomycin will be administered during the study. Patients who require other systemic antibacterial drugs concomitantly are then excluded from the study. Other considerations regarding the inclusion and exclusion criteria are summarized in Table 3.3.1. For example, patients who have history of hypersensitivity to a cephalosporin or penicillin are excluded from the study.

The second example concerns the evaluation of the efficacy of an oral agent for the treatment of patients with noninsulin-dependent diabetes mellitus (NIDDM). As indicated by Cooppan (1994), NIDDM is the most common form of diabetes seen in clinical practice. The prevalence in the United States is about 6.6% and rises to 18% in the elderly. The incidence is about 500,000 new patients every year. NIDDM often have hypertension and hyperlipidemia. In early stages, patients are hyperinsulinemic. Most patients are overweight and have upper body or truncal obesity. The onset of NIDDM is usually above 40 years of age. The disease has a strong genetic basis. It is more frequently seen in native Americans, Mexican Americans, and blacks. The pathophysiology of the disease is due to changes in insulin production and secretion, insulin resistance in liver, muscle, and adipose tissue. A high glucose level could further reduce pancreatic insulin secretion. The treatment for NIDDM patients normally includes (1) diet alone, (2) diet plus oral hypoglycemic drug, and (3) weight control. Note that a mild to moderate weight loss (e.g., 5 to 10 kg) has been shown to improve diabetic control and a moderate calorie restriction (e.g., 250 to 500 calories less than average daily intake) is then recommended for weight control. Based on the above considerations, Table 3.3.2 provides a sample eligibility criteria for the NIDDM study. As can be seen, patients aged 40 years or older—who have a previously established diagnosis of NIDDM and

Table 3.3.2 Eligibility Criteria for a NIDDM Study

<i>A. Inclusion Criteria</i>
1. Males or females aged ≥ 40 years old.
2. Females who are not postmenopausal; they must be nonlactating, incapable of becoming pregnant or of childbearing potential, practicing an effective method of contraception, or have a negative serum pregnancy test documented at screening.
3. Currently suboptimally controlled on diet alone or previously managed on an oral sulfonylurea with an fasting plasma glucose ≥ 126 mg/dL but without symptomatic diabetes.
4. Detectable fasting serum insulin and c-peptide at screening.
5. Normal renal function as defined by serum creatinine of <1.5 mg/dL for men and <1.4 mg/dL for women, and ≤ 1 proteinuria on routine urinalysis.
6. Acceptable liver function as defined by SGOT/AST ≤ 62 U/L and SGPT/ALT ≤ 58 U/L for females and ≤ 90 U/L for males.
<i>B. Exclusion Criteria</i>
1. Markedly symptomatic diabetes, marked polyuria and weight loss $>10\%$.
2. History of hypersensitivity to biguanides.
3. Prior insulin therapy except for acute illness or surgery.
4. Significant cardiovascular disease.
5. Significant renal disease or renal functional impairment as evidenced by a serum creatinine ≥ 1.5 mg/dL for males and ≥ 1.4 mg/dL for females.
6. Significant hepatic disease as evidenced by abnormal liver function as defined as by SGOT/AST > 62 U/L and SGPT/ALT > 58 U/L for females and > 90 U/L for males.
7. Active infectious process such as gangrene and pneumonia.
8. Pulmonary insufficiency.
9. Metabolic acidosis and acute/chronic diabetic ketoacidosis.
10. Any patient for any other condition which, in the investigator's opinion, would make the patient unsuitable for the study or would interfere with the evaluation of the study medication.

are currently controlled on diet alone or were previously managed on an oral sulfonylurea with a fasting plasma glucose greater than 126 mg/dL but without symptomatic diabetes—meet the inclusion criteria for entry. However, exclusion criteria exclude patients who are known to have a history of hypersensitivity of biguanides and significant cardiovascular diseases from the study. Significant cardiovascular diseases may include acute myocardial infarction, unstable angina, congestive heart failure, and arrhythmia.

For the third example, consider a clinical trial comparing the effects of an antidepressant compound to sertraline (Zoloft) on sexual function in patients with previously demonstrated sexual dysfunction with sertraline during treatment for major depression. Segraves (1988, 1992) indicated that patients treated with many psychotropic medications including antidepressants have sexual adverse effects. However, the mechanism by which antidepressants produce sexual dysfunction have not been clearly established. Symptoms of sexual dysfunction may include one or more of the followings: delayed or absent ejaculatory response, partial or total anorgasmia, inadequate lubrication or swelling. The incidence rate for sexual dysfunction for sertraline-treated male patients is 15.5% compared to 2.2% of placebo-treated male patients (Zoloft, 1992). It is believed that both potentiation of peripheral nervous system adrenergic/nonadrenergic activity and increasing brain serotonin (5-HT) level by blocking the neuronal 5-HT reuptake process may induce sexual dysfunction. In order to be eligible for this study, patients must be experiencing sexual dysfunction while being

Table 3.3.3 Eligibility Criteria for Central Nervous System Agents

<i>A. Inclusion Criteria</i>
1. Males or females 18 to 65 years of age. Female patients of childbearing potential must be nonlactating, have a confirmed negative serum pregnancy test prior to enrollment, and be employing an acceptable method of birth control.
2. Patients who are experiencing sexual dysfunction in response to sertraline at a daily dose of 100 mg during their current depressive episode.
3. Treatment with sertraline must have been diagnosed of major depression.
<i>B. Exclusion Criteria</i>
1. Patient having a diagnosis of treatment-resistant depression.
2. History of sexual dysfunction due to any organic condition.
3. Patients who cannot discontinue their current psychotropic medications and/or are likely to require treatment with any prohibited concomitant therapy.
4. History of hypersensitivity to trazodone, etoperidone, or sertraline.
5. Patients receiving any concomitant medication that can produce sexual dysfunction.
6. Patients who have met DSM-IV-R criteria for any significant psychoactive substance use disorder within the 12 months prior to screening.
7. Patients who exhibit a significant risk of committing suicide or have a score ≤ 3 or item 3 "suicide" of the HAM-D scale.
8. Patients who have a significant and/or uncontrolled medical condition.
9. Patients with any clinically significant deviation from normal in the physical or electrocardiographic examinations or medically significant values outside the normal range in clinical laboratory tests.
10. Patients with a positive urine drug screen.
11. Patients with implanted prosthetic devices.
12. Patients who have any other medical condition(s) that can confound the interpretation of the safety and the efficacy data.

treated with sertraline at a daily dose of 100 mg during their current depressive episode. In addition sertraline must have been prescribed for the patient with diagnosis of major depression according to the Diagnostic and Statistical Manual, Fourth Edition (DSM-IV), based on documented patient history. Other considerations for excluding patients from the study including sexual dysfunction due to any organic condition, treatment-resistant depression, or some significant and/or uncontrolled medical conditions. Table 3.3.3 gives a sample list of eligibility criteria for the study. Note that significant and/or uncontrolled conditions may include symptomatic paroxysmal, chronic cardiac arrhythmias, history of stroke, transient ischemic attacks, or history of a positive test for the HIV antibody or antigen.

Patient Selection Process

As discussed above, a set of well-developed eligibility criteria for patient selection can not only best describe the target patient population but also provide a homogeneous sample. The criteria help in reducing bias and variability and consequently increase statistical power of the study. Therefore, in practice, it may be desirable to impose more inclusion and exclusion criteria to further eliminate bias and variability. However, it should be noted that the more criteria that are imposed, the smaller the target patient population will be. Although a smaller patient population may be more homogeneous, it may result in difficulties in

patient recruitment and limitations in the generalization of the findings of the study. Therefore, it is suggested that the considerations not be too restrict to decrease patient enrollment and lose the generality of the target patient population.

In clinical trials, however, the number of patients is usually called for by the study protocol to ensure that the clinical trials can provide valid clinical evaluation of the study medicines with the desired accuracy and precision. It is important then in patient selection to achieve enough patients for the proposed trials. In practice, a single study site may not be feasible for an intended clinical trial due to its limited capacity and resources. Besides, there may not be sufficient patients with the disease available in the area within the scheduled time period of the study. To recruit enough number of patients and to complete the study within the time frame, as an alternative, a multicenter trial is usually considered. If a multicenter trial is to be conducted, the following two questions should be considered:

1. How many study sites should be used?
2. How to select these study sites?

As a rule of thumb, the number of sites should not be greater than the number of patients within each selected study site. This is because statistical comparison between treatments is usually made based on patients (i.e., experimental units) within study sites. It is therefore not desirable to have too many study sites, though it may speed up the enrollment and consequently shorten the completion time of the study. The selection of study sites depends primarily on the following criteria:

1. Individual investigator's qualification and experience for disease.
2. Feasibility of the investigator's site for conducting the proposed trial.
3. Dedication, education, training, and experience of the personnel at the investigator's site.
4. Availability of certain equipments (e.g., magnetic resonance imaging [MRI] or densitometer).
5. Geographic location.

Considerations 1 to 4 ensure that the proposed study will be appropriately carried out in such a way that the differences among investigators is minimized. The geographic location guarantees that the patients enrolled into the study constitute a representative sample from the target patient population. Another important consideration for selection of investigators or study sites is probably their ability to enroll patients and to complete the study within the planned time frame.

For a selected study center, the selection process for patients involves the following concerns:

1. Initial guess of how many patients will meet the eligibility criteria.
2. Screening based on diagnostic criteria.
3. Patient's disease changes.
4. Concurrent diseases/medications.
5. Psychological factors.
6. Informed consent.

At the selected study site, the investigator is often concerned with whether he or she can enroll enough patients for the proposed trial. The investigator usually provides an estimate of how many patients will meet the eligibility criteria based on how many patients he or she has seen at the study site. In practice, such an estimate often overestimates the actual number of patients who will participate the study. Bloomfield (1969) recommends that investigator check the availability and suitability of the patient population at hand through their records or perhaps a formal pilot study.

As indicated by Weintraub and Calimlim (1994), a majority of patients may be excluded at the screening stage of patient selection process due to some administrative reason and the rigor of the diagnostic criteria. Weintraub and Calimlim (1994) point out that administrative reasons such as nonavailability during screening can be as high as 40% in a study intended to evaluate the efficacy of three analgesic treatments and a placebo administered in single doses in double-blind fashion for postoperative pain. Furthermore, in this case 86% were eliminated due to the rigor of diagnostic criteria such as insufficient severity of pain after the operation. Thus, the diagnostic strictures imposed by the clinical trial can decrease the number of available patients even further. It should be noted that small changes in the criteria can make vast differences in patient availability without materially influencing the clinical outcome and its extrapolatability.

It is also recommended that the patient selection process be able to address the issue of specific disease requirements such as disease of a particular severity or duration. For example, a moderate disease status may be preferred because (1) it is realized that patients must be sick enough to get better and (2) patients may be too severe to study. At screening, many patients may be excluded from the study due to disease changes and concurrent diseases requiring concomitant medications. This is true especially for very sick patients who frequently have disease changes and/or concurrent diseases. If we exclude patients who have disease changes and/or concurrent diseases, the patient population under study will become much smaller. Consequently, we may not be able to recruit enough patients for the study. In addition, seasonal factor for some diseases must be taken into account.

Weintraub and Calimlim (1994) point out that ideal participants for clinical trials are patients who will carry through the clinical trials and actively interact with the investigators rather than be passive experimental subjects. It is suggested that psychological factors (e.g., fear of toxicity) be carefully analyzed to enable a patient to make a reasonable judgment about participation in a clinical trial.

At screening prior to the entry of the study, signed and dated written informed consent must be obtained by the investigator from the patient after full disclosure of the potential risks and their nature. Consent must be obtained before a prospective study candidate participates in any study-related procedure, including any change in current therapy required for entry into the study. The fact is that such consent obtained must be recorded in the case report form. In practice, it is not uncommon to allow the investigator to exclude any patient for any condition which, in his/her opinion, would make the patient unsuitable for the study or would interfere with the evaluation of the study medication. For example, the investigator may decide to exclude the patient whose white blood cell count is less than $3500/\text{mm}^3$ or neutrophil count is less than $1500/\text{mm}^3$ from the study of an antidepressant agent comparing with sertraline on sexual function in patients with previously demonstrated sexual dysfunction with sertraline during treatment for major depression as described above. Note that many times the abnormalities that are observed in laboratory tests are due to illnesses unrelated to the disease under study or to other necessary therapeutic interventions.

Ethical Considerations

For many severely destructive diseases such as AIDS, Alzheimer's disease, and cancer, it is unethical to include placebo concurrent control in a clinical trial where an effective alternative remedy is available. It, however, should be noted that the effectiveness and safety of a test agent can only be established by inclusion of a placebo concurrent control. Ethical considerations will definitely affect the patient selection process. In such cases, it is suggested that different numbers of patients be allocated to the treatment arm and the placebo arm in order to reduce the percentage of patients being assigned to the placebo arm. For example, we may consider a two-to-one ratio for a study. In other words, two-thirds of the patients who participate in the study will receive the active treatment and only one-third will receive the placebo. In some comparative clinical trials, if patients are too sick, a certain amount of standard therapy must be permitted for ethical reasons. The use of active control will be discussed further in the next section. Note that it is suggested that placebos be used for trivial, nondangerous, or self-limiting disorders provided that consent is obtained from the patient.

In recent years, ethical considerations for the use of females, children, and the elderly have attracted much attention. For example, in its 1993 revised guidelines for clinical trials, the FDA suggests to include in clinical trials women of childbearing potential who are usually excluded in early drug studies of non-life-threatening disease. Since neonates, infants, and children respond to certain medicines differently than adults, trials of medicines to be used in this age group are always necessary. Therefore, special consideration must be given to the conduct of trials in children. For this purpose, the CPMP (Committee for Proprietary Medicinal Products) of the EC has adopted guidelines on clinical investigation in children. For the geriatric population, it is also important to evaluate any medicine likely to be used in that age group due to the reasons that there is an increasing incidence of adverse events in the elderly and that there are altered pharmacokinetic profiles of some medicines and impaired homeostasis in the elderly.

Note that in the pharmaceutical industry a copy of the final Institution Review Board (IRB) approved informed consent form must be provided to the sponsor before drug supplies will be shipped or enrollment in the study can begin.

3.4 SELECTION OF CONTROLS

In clinical trials, bias and variability can occur in many ways depending on the experimental conditions. These bias and variability will have an impact on the accuracy and reliability of statistical and clinical inference of the trials. Uncontrolled (or noncomparative) studies are rarely of value in clinical research, since definitive efficacy data are unobtainable and data on adverse events can be difficult to interpret. For example, an increase in the incidence of hepatitis during an uncontrolled study may be attributed to the medicine under investigation. Therefore, the FDA requires that adequate well-controlled clinical trials be conducted to provide an unbiased and valid evaluation of the effectiveness and safety of study medicines. The purpose of a well-controlled study is not only to eliminate bias but also to minimize the variability, and consequently to improve the accuracy and reliability of the statistical and clinical inference of the study.

In early 1970s, it was not uncommon for clinical scientists to conduct an uncontrolled clinical trial for scientific evaluation of a therapeutic intervention. Table 3.4.1 summarizes a comparison of positive findings between uncontrolled and controlled clinical trials in

Table 3.4.1 Comparison of the Results Between Uncontrolled and Controlled Trials

Therapeutic Areas	Percent of Positive Findings	
	Uncontrolled	Controlled
Psychiatric (Foulds, 1958)	83%	25%
Antidepressant (Wechsler et al., 1965)	57%	29%
Antidepressant (Smith et al., 1969)	58%	33%
Respiratory distress syndrome (Sinclair, 1966)	89%	50%
Rheumatoid arthritis (O'Brien, 1968)	62%	25%

Source: Summarized and tabulated from Spilker (1991).

selected therapeutic areas. As can be seen from Table 3.4.1, the positive findings of uncontrolled trials were obviously exaggerated. The reason for the overexaggerated positive findings observed in uncontrolled trials, as indicated in the ICH E10 guideline, is that the trials without a control group cannot really allow discrimination of patient outcomes (such as changes in symptoms, signs, or other morbidity) caused by the test drug from the outcomes caused by other confounding factors such as the natural progression of the diseases, observed or patient expectations, or other treatment (ICH E10, 1999). In other words, uncontrolled studies fail to provide an unbiased and reliable clinical/statistical inference regarding what would have happened to patients if they had not received the test drug. Since estimates of the treatment effects are usually extremely biased in the positive direction, the FDA requires that adequate well-controlled studies use a design that permits a valid comparison with a control to provide a quantitative assessment of drug effect (Section 314.126 in Part 21 of CFR). A well-controlled trial is referred to as a trial that is conducted under the experimental conditions such that patient characteristics between treatment groups are homogeneous. One can classify control groups based on two critical attributes: (1) the method of determining who will be in the control group, and (2) the type of treatment received by the patients. The primary methods for assignment of the patients to the control group are either by randomization of the same target patient population or by selection of a control patient population separate from the target patient population treated in the trial. Consequently, the ICH E10 guideline defines a concurrent control group as one chosen from the same population as the test group and treated in a defined way as part of the same trial that studies the test drug (ICH E10, 1999). The test and control groups should be similar with respect to all demographic and baseline characteristics and all on-treatment evaluations and variables in the course of the trial that could influence outcomes or clinical endpoints except for the study treatment. As a result, the difference in outcomes or clinical endpoints observed between the treatment groups at the conclusion of the study is solely due to the different treatments that patients receive rather than other confounding factors as described above. On the other hand, if a control group is chosen separated from the target patient population treated in the same trial, it is referred to as external or historical control. There are in general four types of treatment that a patient can receive in the control group: (1) placebo, (2) no treatment, (3) different dose or regimen of the test drug, and (4) different active treatments. According to Section 314.126 in Part 21 of CFR and the ICH E10 guideline, control group includes placebo concurrent control, dose-response concurrent control, active (positive) concurrent control, no-treatment concurrent control, and external control (historical control), which are described below.

Placebo Concurrent Control

In the past almost two centuries, there has been many heated debates over the use of an inactive placebo group as a reference for evaluation of a test therapy in treatment of patients under aliment or medication conditions. For example, see Brody (1981, 1982), Lundh (1987), Levine (1987), Stanley (1988), and Sanford (1994), Rothman and Michels (1994), Temple and Ellenburg (2000), Ellenburg and Temple (2000), Reynold (2000), and Simon (2000). Brody (1982) defines a placebo as a form of medical therapy, or an intervention designed to simulate medical therapy. A placebo is believed to be without specificity for the condition being treated. A placebo is used either for its symbolic effect or to eliminate observer bias in a controlled experiment. Brody (1982) also indicated that a placebo effect is the change in the patient's condition that is attributable to the symbolic import of the healing intervention rather than to the intervention's specific pharmacologic or physiologic effects. In clinical trials, it is not uncommon to observe the placebo effect. Brody (1982) points out that the placebo effect can be as important as many treatments. Sackett (1989) also indicated that the placebo effect is the most probable cause for symptomatic relief following internal mammary artery ligation experienced by many angia patients. In addition, the influence of the placebo effect is not restricted only to subjective psychological or psychiatric measurements. Placebo also alters objective clinical endpoints such as cholesterol level (Coronary Drug Project Research Group, 1980), laboratory values, and measures of physiologic change (Wolf, 1950), for even the pattern of placebo response resembles the pharmacologic response of the active treatment (Lasagna et al., 1958). For a better understanding of the placebo effect, consider a clinical trial with a 4-week single-blind placebo run-in phase and a 24-week double-blind randomized phase for evaluation of a new agent in three doses with a placebo in treatment of the patients with benign prostatic hyperplasia. The primary clinical endpoints are the proportions of the patients with an at least 3-point improvement of total symptom score or an increase of maximum urinary flow rate greater than 3 mL/s at the end of 24 weeks of the double-blind phase (Boyarsky et al., 1977) which is given in Table 3.4.2. The placebo response rate based on subjective total symptom score is not statistically significantly different from the rates of three doses. Although the placebo effect for the more objective maximum urinary flow rate is less than that based on the symptom score, it is still about 23% and is not statistically significantly different from those of the three doses. The causes of the placebo effect have been speculated for a long time; the effect is now believed to be due to a combination of interactions among patients, physicians, and experimental conditions surrounding the clinical trials (Brody, 1981, 1982; Lundh, 1987; Sanford, 1994). In most

Table 3.4.2 Response Rates by Total Symptom Score and Maximum Urinary Flow Rate (mL/s) at the End of 24 Weeks of the Double-Blind Phase

Dose	Percent of Patients with Improvement of	
	Symptom Score > 3	Maximum Urinary Flow Rate > 3 mL/s
Placebo	41/92(45%)	21/92(23%)
10 mg	36/89(40%)	25/88(28%)
30 mg	38/85(45%)	17/82(21%)
60 mg	36/85(42%)	20/80(25%)

cases, a response or an outcome obtained from a patient receiving an active treatment is a function of four major components, which include (1) the true pharmacological activity of the active ingredient(s), (2) the symptomatic relief provided by the placebo, (3) the natural reversible healing process provided by the body or natural disease progression, and (4) any other known or unknown confounding factors that may have an impact on the response or outcome. The effect contributed by the last three components cannot be unbiasedly estimated unless there is a placebo group in the trials. Therefore the inclusion of a placebo concurrent control in clinical trials is necessary to provide unequivocally and unbiasedly an assessment of the effectiveness and safety of the therapeutic intervention under study.

As indicated earlier, it is unethical to use placebo concurrent controls where symptoms are severe or hazardous and where there exists an alternative therapy with established effectiveness and safety. In practice, it is also unethical to expose patients with severe diseases to study medicines under investigation that may have unknown yet potentially serious even deadly adverse events. The saga of Cardiac Arrhythmia Suppression Trial (Echt et al., 1991) provides a vivid but sad example. If a placebo concurrent group had not been included, neither the excessive risk of death for flecainide and encainide could have been demonstrated nor the assumption of the use of surrogate endpoint ventricular premature contraction (VPC) for mortality could have been proved wrong. Kessler and Feiden (1995) also indicate that the AIDS activists now made an extraordinary plea to the top FDA officials not to approve drugs to treat the disease caused by the human immunodeficiency virus (HIV) too quickly. The reason is that they and their physicians must study in detail the approved antiviral AIDS drugs and examine the efficacy and safety of experimental therapy before they can make an optimal use of these treatments, since many of new experimental drugs were tested without a placebo concurrent group. Spilker (1991) lists the conditions for the ethical use of a placebo concurrent groups in clinical trials. These conditions are summarized in Table 3.4.3. As a result placebo concurrent control should not be selected as the internally controlled group for evaluation of a new treatment if there exists a treatment whose efficacy has already been established for the intended diseases. It is not ethical to use placebo for the care of severe or life-threatening diseases. In all other cases, however, placebo concurrent control should be employed as the standard concurrent control, whenever operationally feasible, for evaluation of the effectiveness and safety of a new therapeutic intervention.

Table 3.4.3 Conditions for the Ethical Inclusion of a Placebo Concurrent Control

1. No standard treatment exists.
2. Standard treatment is ineffective or unproved to be effective.
3. Standard treatment is appropriate for the particular clinical trials.
4. The placebo has been reported to be relatively effective in treating the disease or condition.
5. The disease is mild and lack of treatment is not considered to be medically important.
6. The placebo is given as an add-on treatment to an already existing regimen that is not sufficient to treat patients.
7. Allowing concomitant medicine is one measure of efficacy in these clinical trials.
8. The disease process is characterized by frequent spontaneous exacerbations and remission (e.g., peptic ulcer).
9. "Escape clauses" or points are designed into the protocol.

Source: Spilker (1991).

Examples 3.4.1 and 3.4.2 provide two real examples concerning clinical trials with placebo concurrent control.

Example 3.4.1 Chelation Therapy for Ischemic Heart Disease

Chelation therapy using EDTA is a widely used alternative therapy for ischemic heart disease. In 1993, Grier and Meyers estimated that more than 500,000 people in the United States are treated with EDTA therapy each year. Knudtson et al. (2002) also projected one million U.S. residents will adopt chelation therapy with an annual expenditure of approximately \$400 million U.S. Unfortunately, its efficacy is never fully established. Knudtson et al. and the Program to Assess Alternative Treatment Strategies to Achieve Cardiac Health (PATCH) Investigators (2002) conducted a double-blind, randomized, placebo-controlled trial to determine whether the most commonly used EDTA protocols have a favorable impact on exercise ischemia threshold and quality-of-life measures in patients with stable ischemic heart disease. Random intervention included infusion with either weight-adjusted (40 mg/kg) EDTA chelation therapy or placebo for 3 hours per treatment, twice weekly for 15 weeks and once per month for an additional three months. In addition, patients in both groups also took oral multivitamin therapy.

Example 3.4.2 Evaluation of St John's Wort in Major Depression

St John's wort (*Hypericum perforatum*) is a small flowering weed, and it has been used for a variety of nervous conditions for more than 2000 years. Linde et al. (1996) conducted a meta-analysis of 23 randomized trials of St John's wort extract in 1,757 patients with depressive disorders and concluded that St John's wort was significantly superior to placebo and is an effective agent comparably with standard antidepressant drugs. However, Linde et al. (1996) also indicated that most of the trials used in the meta-analysis have had serious methodological flaws and fail to provide any meaningful interpretation. Therefore, Shelton et al. (2001) conducted a randomized, double-blind, placebo-controlled clinical trial to compare the efficacy and safety of a standardized extract of St John's wort with a placebo in outpatients with major depression. Two hundred adult outpatients diagnosed as having major depression and having a baseline Hamilton Rating Scale for Depression (HAM-D) score of at least 20 were randomized to receive either St John's wort extract (900 mg per day for 4 weeks, increased to 1,200 mg per day in the absence of an adequate response thereafter) or a placebo for 8 weeks.

The results of these two studies provide no evidence to support a better efficacy over placebo in treatment of patients with ischemic heart disease or major depression, respectively. These two examples illustrate that the efficacy and safety of alternative treatments or therapies must be evaluated rigorously using a concurrent placebo control. One of the key elements for the success of a clinical trial using a concurrent placebo control is whether a placebo treatment can be made for matching the active treatment in all aspects such as size, color, coating, texture, taste, or odor. For the chelation therapy for ischemic heart disease, the active treatment is the 500-mL infusion of 5% textrose containing weight adjusted (40 mg/kg) disodium EDTA, with a maximum total dose for each treatment of 3 g, 750 mg of magnesium sulfate, 5 g of ascorbic acid, 5 g of sodium bicarbonate, and 80 mg of Lidocaine. In the placebo infusion, the ETDA was replaced by 20 mL of 0.9% sodium chloride. The resulting infusion solutions were indistinguishable by color and labeling. In addition,

the infusion solution was administrated over 3 hours to minimize the possible unblinding effect of infusion-related adverse events.

Another consideration is whether it is ethical to use a concurrent placebo control in the trials for major depression, which is a serious and potentially life-threatening condition. Therefore, Shelton et al. (2001) considered extensive safeguards. For example, subjects were excluded if they posed a significant risk of suicide at any time during the study. Subjects with a score greater than 2 (i.e., thoughts of death or wishes self dead, but no suicidal ideation or plan) on the item of suicide on the HAM-D were excluded. Subjects with any clinically significant deterioration in their condition from baseline were also excluded. In addition, subjects, who withdrew from the study before the scheduled completion, were immediately offered a standard care as an alternative therapy.

Dose-Response Concurrent Control

As discussed in Chapter 1, the primary objectives for phase II studies are (1) to establish the efficacy, (2) to characterize its dose-response relationship, and (3) to identify the minimum effective and maximum tolerable doses of the therapeutic agent under development. The dose proportionality studies for the assessment of the assumption of linear pharmacokinetics of the test drug usually include at least three doses. Therefore a clinical trial with dose-response concurrent control includes at least two doses of the same test agent. Since the dose-response studies are usually conducted in the phase II stage where the efficacy of the test agent has not yet definitely been established, it is imperative to include a placebo concurrent control to provide an estimate of the absolute efficacy for each dose in addition to the dose-response relationship. The exclusion of placebo concurrent control in a dose-response study could be disastrous and costly. For example, in a major pharmaceutical company, a randomized, double-blind phase II study was conducted to establish the dose-response relationship for a new contrast enhancement agent in conjunction with MRI in diagnosis of malignant liver tumors in patients with known focal liver lesions. Despite suggestions of changes by the project statistician because of (1) the exclusion of a placebo concurrent control, (2) the use of an invalidated scale for diagnostic confidence, and (3) visual evaluation of the pre- and postcontrast films as the primary endpoint, the trial was conducted as planned without any of the modifications. Table 3.4.4 provides the proportion of the patients with good or excellent improvement for diagnostic confidence or visual evaluation. As can be seen from Table 3.4.4, there was no dose-response at all. It should be noted that without a placebo concurrent control, it is impossible to assess whether a response rate between 60 to 65% observed in this trial really demonstrates the true efficacy of the test agent because there were no trials with a placebo concurrent control ever conducted.

Examples 3.4.3 and 3.4.4 provide two real examples concerning clinical trials using dose-response concurrent control.

Table 3.4.4 Percent of Patients with Good or Excellent Improvement in Diagnostic Confidence and Visual Evaluation

Dose	Percent of Patients with Good or Excellent Improvement	
	Diagnostic Confidence	Visual Evaluation
12.5 μmol	45/77 (58%)	51/79 (65%)
25.0 μmol	47/78 (60%)	52/82 (63%)
50.0 μmol	45/76 (59%)	50/78 (64%)

Example 3.4.3 Treatment of Erectile Dysfunction

Erectile dysfunction is estimated to affect up to 30 million men in the United States. Effective oral therapy for treatment of the males with erectile dysfunction has been sought for a long time. Sildenafil is considered a promising agent for erectile dysfunction. However, it is recognized that sildenafil is a potent inhibitor of cyclic guanosine monophosphate in the corpus cavernosum. It increases the penile response to sexual stimulation. Goldstein et al. (1998) conducted a randomized, double-blind study to investigate the dose response of the efficacy and safety of an oral sildenafil in men with erectile dysfunction of organic, psychogenic, or mixed causes. In this fixed-dose study, 532 men were randomly assigned to receive 100, 50, and 25 mg of sildenafil or a placebo, approximately one hour before planned sexual activity but not more than once daily, for 24 weeks. This dose-response study used three doses of the active treatment in the multiples of 25 mg (i.e., 1, 2, and 4) and a placebo concurrent control. Hence, the doses are equally spaced in logarithmic scale (based 2).

Example 3.4.4 Dose-Response of Smallpox Vaccine

The World Health Assembly (WHA) declared that the world is free of smallpox (vaccinia virus) in 1982, and the United States ended its general use of the smallpox vaccine in 1972. As a result, less than half of the world's population has been exposed either to smallpox or to the vaccine. Consequently, several governments or world health authorities have warned that smallpox is a potential biological weapon with a serious threat, which can cause a catastrophic effect in an unimmunized population. On the other hand, the last lot of vaccinia vaccine was manufactured in the United States in 1982. Frey et al. (2002) conducted a randomized, double-blind trial using three dilutions of vaccinia virus vaccine in previous unimmunized adults to assess the clinical success rates, humoral response, and virus-specific activity of cytotoxic T cells and interferon- γ -producing T cells. Three doses of the vaccinia virus vaccine include undiluted vaccine ($10^{7.8}$ plaque-forming units [pfu] per milliliter), 1 : 10 dilution ($10^{6.5}$ pfu per milliliter), and 1 : 100 dilution ($10^{5.0}$ pfu per milliliter). These doses are also equally spaced in logarithmic scale (based 10).

Active (Positive) Concurrent Control

During the development of a new test agent, it may be of interest to establish a superior efficacy than the standard agent or to show therapeutic equivalence in efficacy to the standard therapy but with a better safety profile. For these purposes clinical trials are usually conducted with active agents concurrently. In many cases active treatments are employed for ethical reasons. If the trials are designed to serve as adequate well-controlled trials for providing substantial evidence of efficacy and safety for drug approval, the active treatment concurrent control must unequivocally demonstrate its superior efficacy in pivotal trials with a placebo concurrent control. Otherwise the trials must include a placebo concurrent placebo in addition to the active treatment concurrent control. In some cases clinical trials are conducted to establish therapeutic equivalence to a standard therapy because of no systematic absorption of a different route of administration such as the metered dose inhaler (MDI) for asthma and retin-A for acne, or because of inadequacy of pharmacokinetic measures for chemicals such as sucralfate for acute duodenal ulcer (Liu and Chow, 1993; Liu, 1995a). However, as indicated by Temple (1982) and Huque and Dubey (1990), equivalence between two active agents demonstrated in an active control trial can imply that both agents are efficacious or

both are ineffectual. Therefore it is important to always include a placebo concurrent control in the active control trials unless a superior efficacy has been established and accepted by the regulatory authority. For additional references of controversial issues regarding equivalence/noninferiority trials, see Ware and Antman (1997) and Djulbegovic and Clarke (2001). Note that active control trials will be discussed further in Chapter 7.

The following examples (Examples 3.4.5 and 3.4.6) provide two real examples concerning clinical trials using active concurrent control.

Example 3.4.5 Continuous Infusion Versus Double-Bolus Administration of AMI

Accelerated infusion of alteplase (tissue plasminogen activator) over a period of 90 minutes produced the lowest mortality rates in GUSTO I trial (1993). On the other hand, double-bolus administration of alteplase in two bolus doses given 30 minutes apart further shortens the duration of administration. The Continuous Infusion versus Double-bolus Administration of Ateplase (COBALT) investigators (1997) conducted a randomized trial to test the hypothesis that double-bolus alteplase is at least as effective as accelerated infusion. The test treatment for the COBALT was the administration of a bolus of 50 mg of alteplase over a period of 1 to 3 minutes followed 30 minutes later by a second bolus of 50 mg (or 40 mg for patients with a body weight less than 60 kg). The active concurrent control was an intravenous bolus of 15 mg followed by an infusion of 0.75 mg per kilogram of body weight over a 30-minute period, not to exceed 50 mg, and then by an infusion of 0.5 mg per kilogram, up to a total of 35 mg, for 60 minutes. This trial is a typical example of an noninferiority trial to test the hypothesis that the efficacy of a shorter and easier administration of double-bolus of alteplase is no worse than that of the standard administration of accelerated infusion.

Example 3.4.6 Maintenance Antiretroviral Therapies in HIV Infected Subjects

Three-drug antiretroviral therapy with zidovudine, lamivudine, and indinavir can suppress the level of human immunodeficiency virus (HIV) RNA in plasma below the threshold of detection for two years or more. Havlir et al. and AIDS Clinical Trial Group Study 343 team (1998) conducted a trial to investigate whether a less-intensive maintenance regimen could sustain viral suppression after an initial response to combination therapy. This study consisted of two phases. The induction phase was a period of 24 weeks in which the HIV-infected subjects with CD4 cell counts greater than 200 per cubic millimeter received open-label treatment with indinavir 800 mg, t.i.d., lamivudine, 150 mg, b.i.d., and zidovudine, 300 mg, t.i.d. For the second part of the study, subjects who had less than 200 copies of HIV RNA per milliliter of plasma after 16, 20, and 24 weeks of induction therapy were randomly assigned to receive one of three treatments in a double-blind fashion. The treatments included the original three-drug therapy, indinavir monotherapy, and two-drug combination of zidovudine and lamivudine. For this study, the test treatments were indinavir monotherapy and two-drug combination of zidovudine and lamivudine, while the standard three-drug therapy served as the active concurrent control.

No Treatment Concurrent Control

For certain diseases, under the assumptions that (1) the objective measurements for effectiveness are available and can be obtained in a very short period of time and (2) the placebo effect is negligible, the test agent can be compared concurrently with no treatment. In these

cases the FDA requires that patients be randomized to receive either the test agent or the no treatment concurrent control (Section 314.122 in Part 21 of CFR). In practice, however, it is recommended that no treatment concurrent controls should be avoided if possible during clinical development of phases I–III trials of new agents due to the reasons that it is not good clinical practice and that it fails to simulate the psychological effect of the placebo on efficacy.

Examples 3.4.7 and 3.4.8 provide two real examples of clinical trials using no-treatment concurrent control.

Example 3.4.7 Treatment of Primary Pulmonary Hypertension

Primary pulmonary hypertension is a serious and progressive disease for which few treatments have been shown in a prospective, randomized trial to improve survival. The Primary Pulmonary Hypertensive Study Group (Barst et al., 1996) conducted a prospective, randomized, multicenter open-label trial to compare the effects of continuous IV infusion of epoprostenol plus conventional therapy with those of conventional therapy alone in 81 patients with severe primary pulmonary hypertension (New York Heart Association functional class III or IV). For this study, the test treatment is epoprostenol plus conventional therapy while the control treatment is no-treatment concurrent control plus conventional therapy.

Example 3.4.8 The First Controlled Clinical Trial

As reported by Boylston (2002), in 1767, the leading cause of death among children in the city of London, the United Kingdom, was smallpox. The infection was epidemic and killed one in four children born in the city. At that time, Dr. William Watson was the physician for the Foundling Hospital. Because of the smallpox epidemic, the governor of the hospital ordered that all children who were not already immune to smallpox be inoculated. However, there were two fundamental questions of inoculation: What was the best source of the inoculum? Did mercury (a populated component of pretreatment therapy at that time) provide any clinical benefit? Dr. Watson designed a group of three trials to explore both questions. He recognized that he needed to study a large number of children of similar age in both genders. On October 12, 1767, Dr. Watson conducted his first trial. Thirty-one children were divided into three groups. The source of inoculum was the early lesion from a patient with naturally acquired smallpox. The first group of 10 children (5 of each gender) received a mixture of mercury and jalap (a laxative) before and after the puncture, the second group of 10 children (5 of each gender) received senna and syrup of roses (a mild laxative) on three occasions, and the last group of 11 boys received no treatment for pretreatment regimen. He also tried to keep all other known confounding factors as similar as possible for all children in the trials. For example, all children had the same diet, wore similar clothes, played in the same field, slept in the same dormitory, and were inoculated at the same time and place with the same material. Therefore, the only difference was the treatment children received. Dr. Watson was the first clinician who introduced a no-treatment concurrent control in a clinical trial. The second trial had a similar pretreatment regimen also with a no-treatment concurrent control with the mature lesions from inoculated patients as the source of inoculum. The last trial did not have pretreatment regimen. All 20 children were inoculated using late lesion from inoculated patients as the source of inoculum. More details regarding the trials can be found in Boylston (2002).

Historical Control

In clinical research sometimes it is of interest to compare the results of the test treatment with those of other active treatments or the historical experience of a disease or condition that is adequately documented. Basically, historical data are obtained in two ways. One is from the same group of patients who received no treatment, the same treatment, or different treatments at different times. The other is from different patients who received no treatment, the same treatment, or different treatments at different times. In either case the data of historical control are not obtained concurrently. Therefore, the experimental conditions of the trials are not obtained concurrently for both the test and control groups. Hence, Section 314.122 of Part 21 of CFR indicates that the historical control are reserved for the special diseases with high and predictable mortality such as certain malignant cancers or for the agents in which the effect of the drug is self-evident such as general anesthetics.

In summary, for clinical development of phases I–III trials of a new test agent, the principle of good clinical practice for regulatory approval is to dictate the placebo concurrent control as the fundamental referenced control for unbiased evaluation of effectiveness and safety unless unequivocal evidence proves that it is unnecessary.

3.5 STATISTICAL CONSIDERATIONS

At the planning stage some statistical considerations regarding the manner in which the data will be tabulated and analyzed at the end of the study should be carefully considered. These considerations include the primary and secondary response variables, the criteria for efficacy and safety assessment, sample size estimation, possible interim analysis and data monitoring, and statistical and clinical inference. We will now describe these considerations.

Efficacy and Safety Assessment

For a clinical trial, it is recognized that it is impossible to address all questions with one trial. Therefore, it is important to identify the primary and secondary response variables that will be used to address the scientific and/or medical questions of interest. The response variables (or clinical endpoints) are usually chosen at the outset, since they are needed to fulfill the study objectives. Once the response variables are chosen, the possible outcomes of treatment are defined, and those showing efficacy and safety are clearly indicated. In practice, it is suggested that the selected clinical endpoints be validated (reliable and reproducible), widely available, understandable, and accepted. For example, in an antibiotic trial the outcome might be defined as cure, cure with relapse, or treatment failure, and the response variables may be pyrexia, dysuria, and frequency of urination. The criteria for the evaluation of a cure could be that all signs or symptoms of urinary tract infection are resolved during the study period. For another example, in an antihypertensive trial the outcome of treatment might be defined as normalization, partial response, or failure, and the response variable would be change in blood pressure. The criteria for normalization and partial response could be that diastolic pressure is less than 90 mmHg and that diastolic blood pressure is reduced by more than 10% from baseline, respectively.

For efficacy assessment, once the primary efficacy variable is identified, the criteria for the evaluability of the patients should be precisely defined. For example, we may conduct an analysis based on all patients with any effectiveness observation or with a certain minimum number of observations. In some cases clinical scientists may be interested in

analyzing patients who complete the trial (or completer analysis) or all patients with an observation during a particular time window. To provide a fair assessment of efficacy, sometimes it may be of interest to analyze only patients with a specified degree of compliance, such as patients who took 80% to 120% of the doses during the course of the trial. It should be noted that the evaluability criteria should be clearly defined in the study protocol. As indicated in the FDA guidelines, although a reduced subset of the patients is usually preferred for the data analysis, it is recommended that an additional intent-to-treat analysis using all randomized patients be performed.

For safety evaluation the FDA requires that all patients entered into treatment who received at least one dose of the treatment must be included in the safety analysis. Safety evaluation is usually performed based on clinical and laboratory tests. To provide an effective evaluation, it is suggested that the following should be provided:

1. Parameters to be measured.
2. Timing and frequency.
3. Normal values for laboratory parameters.
4. Definition of test abnormalities.

The primary safety variable is the incidence of adverse event, which is defined as any illness, sign, or symptom that has appeared or worsened during the course of the clinical study regardless of causal relationship to the medicine under study. The FDA suggests that basic display of adverse event rates be used to compare rates in treatment and control groups. In addition, if the study size permits, the more common adverse events that seem to be drug related should be examined for their relationship to dosage and to mg/kg dose, to dose regimen, to duration of treatment, to total dose, to demographic characteristics, or to other baseline features if data are available. However, the FDA also points out that it is not intended that every adverse be subjected to rigorous statistical evaluation.

Sample Size Estimation

For assessment of the effectiveness and safety of a study drug, a typical approach is first to show that the study drug is statistically significant from a placebo control. If there is a statistically significant difference, we then demonstrate that the trial has a high probability of correctly detecting a clinically meaningful difference. The probability of correctly detecting a clinically meaningful difference is known as the (statistical) power of the trial. In clinical trials, for a given significance level, we can increase the statistical power by increasing the sample size. In practice, a pre-study power analysis for sample size estimation is usually performed to ensure that the intended trials have a desired power (e.g., 80%) for addressing the scientific/medical questions of interest. In clinical trials, we can classify sample size estimation as either sample size determination or sample size justification. The purpose of a sample size determination is to find an appropriate sample size based on the information (the desired power, variability and clinically meaningful differences, etc.) provided by clinical scientists. If the sample size has been chosen based on medical/marketing considerations, then it is necessary to provide a sample size justification for the chosen sample size such as "What difference can be detected with the desired power for the chosen sample size?" It should be noted that a larger sample size will allow us to detect a smaller difference if the difference indeed exists.

Table 3.5.1 Sample Size Determination

Power	Standard Deviation (%)	Sample Size ^a	
		Clinical 10%	Difference 15%
80%	10	32	14
	20	126	56
	30	284	126
90%	10	44	20
	20	170	76
	30	380	170

^aSample sizes were obtained based on a two-sided test for two independent samples at the $\alpha = 5\%$ level of significance.

For sample size determination, Table 3.5.1 provides some examples of required sample sizes for achieving desired power to detect some clinically meaningful differences under the assumption of various standard deviations. The estimated sample sizes were obtained based on a two-sided test for two independent samples with the 5% level of significance. For example, a total of 32 patients is needed to have a 80% power for detection of one standard deviation difference. Additional 12 patients are required to increase the power from 80% to 90%. As can be seen from Table 3.5.1, the sample size increases as the standard deviation increases. In addition a larger sample size is required to detect a smaller difference. On the other hand, Table 3.5.2 gives statistical justifications for differences that can be detected for some chosen sample sizes. For example, the selected sample size of 100 will have a 80% power for detection of an approximately half standard deviation difference. Note that the difference that can be detected based on the selected sample size may not be of clinically meaningful difference.

It should be noted that the sample size determination/justification should be carried out based on appropriate statistics under the selected design. Different study designs and testing hypotheses may result in different sample size requirements for achieving a desired

Table 3.5.2 Sample Size Justification

Sample Size	Standard Deviation (%)	Detected Difference ^a	
		80%	90%
100	10	5.6	6.5
	20	11.2	13.0
	30	16.8	19.5
200	10	4.0	4.6
	20	8.0	9.2
	30	12.0	13.8

^aThe numbers were obtained based on a two-sided test for two independent samples at the $\alpha = 5\%$ level of significance.

power. Therefore it is recommended that the following be considered when performing a pre-study power analysis for sample size estimation:

1. What design is to be used?
2. What hypotheses are to be tested?
3. What statistic is to be performed?

If the selected design is a parallel design and/or a two-sided test is used, it may require more patients to reach the desired power. Under a selected design, sample size requirements are different for testing point hypotheses and interval hypotheses. As discussed in the previous chapter, an interval hypotheses is intended for establishment of clinical equivalence. The FDA indicates that for a positive control study intended to show that a new therapy is at least as effective as the standard therapy, the sample size determination should specify a clinically meaningful difference indicating that the new therapy is clinically equivalent if the difference is smaller than such a difference. The power to detect a treatment difference should be given.

Interim Analysis and Data Monitoring

Interim analysis and data monitoring is a process of examining and/or analyzing data accumulating in a clinical trial, either formally or informally, during the conduct of the clinical trial. The nature and intent of data monitoring and interim analysis in the pharmaceutical industry are often misunderstood. As indicated by the Biostatistics and Medical Ad Hoc Committee (BMAHC) on Interim Analysis of the Pharmaceutical Manufacturing Association (PMA), the following three issues should be addressed when planning an interim analysis (PMA, 1989):

- Protection of the overall type I error rate in formal confirmatory clinical trials designed to establish efficacy.
- Safeguarding of the blinding of a study.
- Use of interim analyses for administrative or planning purposes to generate hypotheses for future studies or to assess safety.

For the protection of the overall type I error, there are many methods available in the literature. For example, see Pocock (1977), O'Brien and Fleming (1979), Peto et al. (1976), Slud and Wei (1982), Lan and DeMets (1983), Lan and Wittes (1988), and Jennison and Turnball (2000). The protection of the overall type I error can usually be achieved through a carefully planned study protocol. The safeguarding of the blind is a critical issue that has a great impact on the credibility of the study. Therefore it is suggested that the sponsors to develop formal procedures to ensure that the dissemination of the results of interim analyses is controlled in such a way as to minimize the potential bias. The third issue considers an interim analysis as a study management tool for addressing some important questions during the conduct of the study. Such analyses, which are known as administrative look, are usually performed on an unblinded basis and without adjustment of *p*-values.

Interim analysis and/or data monitoring provides an administrative tool for terminating a trial during which is observed either a superior efficacy or an excessive safety risk in the treatments presented to patients. Currently all clinical trials sponsored by NIH are required to

perform interim analyses. If no interim analyses is intended, the reasons why an interim analysis is not necessary for the study are to be clearly stated in the study protocol. Since the late 1980s pharmaceutical companies have begun to recognize the need of interim analyses and consequently have started to perform interim analyses and administrative data monitoring for their sponsored trials in an expeditious manner. Since blinding plays an important role in protecting the integrity of clinical trials, the BMAHC was formed to examine the impact of interim analysis on blindness (PMA, 1989, 1993). In their position paper (PMA, 1993), the BMAHC emphasized that blinding (masking) is an important issue, since the interim analysis requires that the study be unblinded. Knowledge of early trends or lack thereof can bias the remainder of the study and result in changes in the patient recruitment. Knowledge of treatment assigned to individual patients can also introduce bias and serious dropouts on the validity of results. The committee suggested that an SOP be developed to describe

- Who will have access to the randomization codes?
- How the blinding will be broken?
- Who will have access to the interim results?
- Whether ongoing patients will be included in the analysis?

Williams et al. (1993) point out that major pharmaceutical companies such as Merck have developed their own internal SOPs for triple-blind policy for all phase III and IV studies and most of phase II trials. Moreover, they state that interim evaluation should be performed by a party that is not involved in the actual conduct of the study. The results of the interim analysis should not be provided in any form to those individuals involved in the conduct of the trial in order to avoid any temptation to alter the study design and to introduce any potential bias. They suggest that the following procedures be imposed to ensure the blindness when performing an interim analysis.

- Merge of randomization codes with patient identification numbers for the use of interim analysis must be performed by a low-level statistician who is not directly involved in the study.
- Identity of treatments received by individuals must not be known.
- Only the minimum information required to meet the objectives of the interim analysis can be presented and only to the few individuals who are responsible for decision making on the drug's development.
- Detailed procedures for the implementation of interim analysis such as unblindedness, decision making, and the frequency of interim analyses must be fully documented and available for external review.

Since interim analyses require not only that the randomization codes be unblinded during the clinical trials but also that the results be disseminated to either the external or internal data-monitoring board, blindness is in fact compromised to some extent and a bias will always be introduced. The FDA has expressed some concerns regarding the issue of only an internal data-monitoring board. O'Neill (1993) indicates that: FDA is primarily concerned that a study can be biased by monitoring practices and procedures, and since the monitoring group has a vested interest in the product being evaluated, the study might be compromised to the extent that it will not support the scientific regulatory standards for drug approval.

Therefore, in addition to the procedures recommended by Merck, we also suggest the following:

- The external data-monitoring board should include clinical and statistical experts from academic institutions in the therapeutic areas under investigation.
- The interim analysis should be performed by the statistical members of the data-monitoring board, using the database merged by the low-level statistician, who is not involved with the trial, from the randomization codes and patient identification numbers with pre-specified efficacy and safety endpoints for the interim analysis in the protocol.
- The external data-monitoring board should have the authority to make decisions regarding when, how, what, and to whom (including those in the top management who make the decision on the drug's development) the results of the interim analysis should be available.

Note that the FDA guideline declares that all interim analyses, formal or informal, by any study participant, sponsor staff member, or data-monitoring group must be described in full even if the treatment groups are not identified. The need for statistical adjustment because of such analyses should be addressed. More details regarding interim analysis will be given later in this book.

Statistical and Clinical Inference

Statistical and clinical inferences are usually drawn based on clinical data collected from controlled randomized trials. Statistical and clinical inferences are derived from statistical tests under the assumption that the selected sample (i.e., a group of patients) is a random sample from the targeted patient population. A random sample is referred to a representative sample. However, in most clinical trials patients are not selected from the target population in a random fashion. In practice, for a clinical trial we usually select study sites (or centers) first. Patients are then recruited at each selected study site to form a sample for the intended clinical trial. Thus no formal sampling theory can be applied to derive a valid statistical inference regarding the target patient population. Consequently the clinician cannot draw a valid statistical inference to clinical practice.

It should be noted that statistical inference is only a part of induction process for the conclusions obtained from clinical trials, and it should not preclude the possibility of a meaningful clinical inference. If the inclusion and exclusion criteria are precisely stated in the study protocol before the trial is conducted, then the demographic characteristics at baseline of the patients can be used to describe the patient population from which the sample of the patients in the trial is drawn. As a result the population model described earlier can be invoked to provide a basis for a clinical inference about the patient population. This concept of clinical inference as a form of induction from sample to population is based on external validity. For example, suppose that the inclusion criteria of a single-center clinical trial allow only enrolling patients within a very narrow age range. Also suppose that there is another study with the same sample size that is a multicenter trial for investigation of the same drug. However, the second study has a much wider age range. The second study is more appropriate to make clinical inference externally simply because it is a multicenter trial with a wider age range than the first. Figure 3.5.1 provides a diagram of statistical and clinical inference with respect to the relationship among randomization for selection and assignment of patients and internal and external validity.

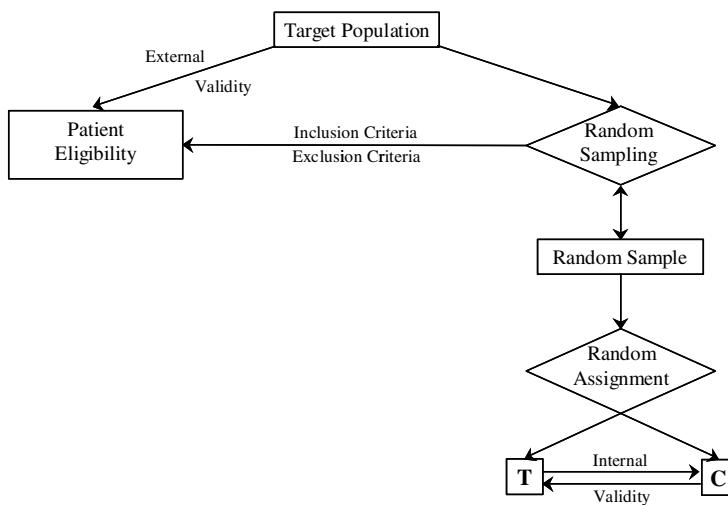


Figure 3.5.1 Diagram of statistical and clinical inference.

Rubins (1994) indicated that any single trial is unlikely to have a major impact on a physician's medical practice. Clinical inference is rarely based on the results of a single clinical trial. In order to investigate the effectiveness and safety of a therapeutic agent or a class of therapeutic agents, a series of clinical trials with the same design and concurrent control is usually conducted over patients with similar but different characteristics. Note that these clinical trials may be conducted by different investigators at different study centers in different countries. Petitti (1994) points out that the technique of the meta-analysis may provide the most conclusive evidence of clinical inference of the therapeutic agents for medical practice because of its external validity. For example, Yusuf et al. (1985) conducted a meta-analysis based on 64 randomized beta-blocker trials to conclude, once for all, the benefit of long-term use of the beta-blocker after myocardial infarction in reduction of mortality. Yusuf et al. (1985) report that there exists no important difference in benefit among different beta-blockers. On the other hand, in the CAST trial, the results indicated that patients who received class IC antiarrhythmic agents (e.g., flecainide or encainide) after myocardial infarction had three times as high as mortality rate than those patients who received the placebo. Since the CAST trial is a single but relatively large trial, the finding was confirmed through a meta-analysis combining 10 clinical trials performed by Hines et al. (1989) in order to prove the harmful effect of the use of class IC arrhythmic agents after myocardial infarction.

In most clinical trial, women with childbearing potential are excluded from the studies. This exclusion limits the generalization of the findings to the larger population. In 1993 the FDA revised its 1977 guidelines to include women of childbearing potential who were previously excluded in early drug studies of non-life-threatening diseases. This revised guideline, however, does not apply to bioequivalence trials which are required by the FDA for approval of generic drug products. As indicated by Henderson (1993), bioequivalence trials are mostly conducted in normal male volunteers aged between 18 and 50 years old whose weights are within 10% of their ideal body weight. As a result of this exclusion of female, patients, and elderly from the studies, bioequivalence trials have very limited external validity. However, under the fundamental bioequivalence assumption, clinical

inference based on rather limited bioequivalence studies is still acceptable to the FDA. The fundamental bioequivalence assumption states that if two drug products are bioequivalent, then they will reach the same therapeutic effects or they are therapeutically equivalent. The legal basis of this fundamental bioequivalence assumption is from the *Drug Price Competition and Patent Term Restoration Act* passed by the United States Congress in 1984. The FDA allows a generic drug to be used as a substitute of a brand-name drug if the generic drug product is shown to be bioequivalent to the brand-name drug. When a brand-name drug is off patent, it is expected there might be a number of generic copies for the brand-name drug approved by the FDA. Although each generic copy of the brand-name drug can be used as a substitute for the brand-name drug, the FDA does not indicate that these generic copies of the same brand-name drug can be used interchangeably. It is therefore important to investigate the overall bioequivalence and inconsistencies among all generic copies of the same brand-name. For this purpose, Chow and Liu (1997) proposed the concept of meta-analysis for postapproval bioequivalence review.

3.6 OTHER ISSUES

In addition to the basic design considerations described above, some specific considerations in planning a design for a clinical trial are given below.

Single Site versus Multi-sites

In most clinical trials, multistudy sites (multicenter) are considered. This is often due to a limitation in the capacity of a single site. Besides, multistudy sites can increase patient enrollment and consequently shorten the duration of the study. Although a multicenter study has its advantages, it also suffers from some difficulties. For example, if the enrollment is too slow, the sponsor may wish to (1) terminate the inefficient study sites, (2) increase enrollments for the most aggressive sites, or (3) open new sites during the course of the trial. Each action may introduce potential biases to the study. In addition, the sponsor may ship unused portions of the study drugs from the terminated sites to the newly opened sites. This can increase the chance of mixing up the randomization schedules and consequently decrease the reliability of the study. For a multicenter study, the FDA requires that the treatment-by-center interaction be carefully investigated. No overall conclusion can be made across study sites if treatment-by-study-site interaction is present. In practice, it is desirable to have fewer study sites than the number of patients within each site. This is because the comparison is usually made between patients within sites. More centers may increase the chance of observing the treatment-by-study-site interaction. In addition different centers may have different standards for clinical evaluation and laboratory tests. It is therefore suggested that a central laboratory be employed whenever possible to provide appropriate and consistent measures.

Treatment Duration

In clinical trials it is desirable to collect clinical data, perform data analysis, and draw statistical inferences on the scientific questions that these clinical trials intend to address in a timely fashion. The duration of a clinical trial depends on the half-life of the study drug, the disease status, and the intended indication. For example, a 10- to 14-day treatment is

usually required for antibiotic agents in order to reach the optimal therapeutic effect. For chronic diseases a longer treatment is necessary to observe significant improvements. If the intended indication is for bone loss, then the duration can be even longer. A typical clinical trial usually consists of three phases: a placebo run-in phase, an active treatment period, and a follow-up or maintenance phase. A placebo run-in phase is usually considered in order to remove effects of previous therapy or to stabilize patients' conditions prior to randomization. An active treatment period is to determine whether the study drug achieves the desired efficacy, and a follow-up or maintenance phase is to monitor the safety of the study drug.

Treatment duration is an important consideration in designing clinical trials. It has an impact on the evaluation of the study drug. An inadequate duration of treatment may not provide an unbiased and valid assessment of the true response rate of the study drug. It may also result in a high dropout rate due to the lack of efficacy. For example, the response rates at different time points may be different for antibiotic agents. It is therefore suggested that an adequate duration of treatment be considered in clinical trials to provide an unbiased and valid assessment of the study drug.

Patient Compliance

It has long been realized that patients may not follow instructions for the use of medication. As indicated by Cramer et al. (1989), overdosing, underdosing, and erratic dosing commonly occur in all patient populations regardless of the severity of illness. Patient compliance has an impact on the evaluability of patients for clinical evaluation of efficacy and safety of the study medicine. Stewart and Cluff (1972) indicate that the extent of patient default is between 20% and 80%. A typical approach to monitoring patient compliance is pill counting. Cramer et al. (1989) suggest using a medication event monitor system (MEMS) to study patients' pill-taking habits. They found that even for a once daily dose schedule, there is only 87% compliance. For the four times daily regimen, the compliance rate drops sharply to 39%. They estimated that the overall compliance rate is around 76%. In practice, to ensure that patients are eligible for clinical evaluation, a set of criteria is usually imposed on patient compliance. For example, patients are considered evaluable if they are within 80% and 120% of compliance.

Wang, Hsuan, and Chow (1996) classify the concept of compliance into compliance and adherence. A patient is said to have a poor compliance if he or she fails to take the drug at the prescribed dose. A poor adherence is referred to as failure to take the drug at the scheduled times. Poor compliance may result in treatment failure and possible adverse reactions. It has been observed that 5% of patients have a drug-induced disease on admission to the hospital (e.g., see Seidl et al., 1996; Hurwitz, 1969). Seidl et al. (1966) also indicate that adverse drug reactions are the seventh most common cause for admission to hospital. Elaboration of other medical consequences of noncompliance can be found in Glanz et al. (1984) and Cramer et al. (1989).

Missing Value and Dropout

Another issue that is worthy of attention is possible dropouts or missing values. Dropouts can be related to the duration, the nature of the disease, and the effectiveness and toxicity of the study drug. It may be misleading to ignore the patients who dropped out prior to the maturity of the study. For example, if more dropouts occur in one treatment group, a bias may have been introduced to the trial. In practice, more patients are usually enrolled to

account for possible dropouts so that the study will have sufficient evaluable patients to achieve the desired power. When there are a large number of dropouts, it is suggested that the causes of the dropouts be carefully evaluated. It can be a great concern if dropouts are related to the ineffectiveness and side effects of the study drug.

As was mentioned earlier, a clinical trial is more likely to be a multicenter trial. Matts and Lachin (1988) indicate that an adequate statistical analysis is a stratified analysis with study center as a stratum. If study center is not considered a stratum and omitted in the analysis, then such an unstratified analysis will likely produce a conservative test. In many clinical trials unplanned post hoc subgroup analyses based on some covariates (patient characteristics) are required to answer some important clinical questions, even though a stratified randomization with respect to the covariates is not performed. If the covariates used for the classification of subgroups are statistically independent of the random assignment of patients to the treatments, then the stratified analysis based on the covariates will be a valid statistical test. A typical approach is to perform an analysis of covariance (ANCOVA), which can include the following as covariates: (1) demographic factors such as age, gender, and race, (2) geographical region such as study center, and (3) some baseline characteristics such as disease severity at entry (Snedecor and Cochran, 1980). In clinical trials, it is almost impossible to collect data from all patients in order to cover all of the information regarding patient characteristics of interest. It is therefore not uncommon to have missing data in some combinations of covariates in clinical trials. One way to handle this problem of missing data is to perform the analyses only on the set of patients with the complete data. This approach is in fact a post hoc stratified analysis based on a covariate that indicates whether a patient has complete data or not. If the missing mechanism is independent of the random assignment of patients to the treatment, then the resulting analyses based only on the subset of the patients with complete data will be statistically valid. Rubin (1976) and Little and Rubin (1987) refer to the assumption of independence between the missing mechanism and treatment assignment as missing at random. However, this assumption cannot be verified or tested. Although the analyses might be valid, they are inefficient because they are based on the subset of the patients with complete data.

3.7 DISCUSSION

In practice, it should be realized that clinical trials are conducted *in* humans and are done *by* humans. Therefore, some issues need to be seriously considered when planning a clinical trial. These issues include safety, compliance, and human error/bias. For example, is it safe or ethical to conduct placebo-controlled antibiotic studies? Patient compliance depends on the cooperation of the patient. It depends on the duration of the study, the duration of visits, the frequency of visits, and perhaps the timing of visits. Human error/bias can be controlled through placebo control, blinding, lead-in, and education. Note that controlled studies are usually referred to as those in which the test treatment is compared with a control. For demonstration of the effectiveness and safety of a test treatment, uncontrolled studies are rarely of value in clinical research, since definitive efficacy data are unobtainable and data on adverse events may be difficult to interpret. As a result, to answer the question, in addition to the definition of patient population and the selection of clinical endpoints, controlled studies are the best way to factor out human error/bias.

Besides the basic design issues described above, clinical data quality assurance is also an important consideration when planning a clinical trial. The success of a clinical trial

depends on the quality of the collected clinical data. The quality of clinical data depends on the case report forms used to capture the information. Inefficient case report forms can be disastrous to a study. Therefore effectively designed case report forms are necessary to ensure the quality of clinical data and consequently to ensure the success of the study. It is recommended that a biostatistician be involved in the design and review of the case report forms to ensure that these forms capture all of the relevant information for data analysis.

4

RANDOMIZATION AND BLINDING

4.1 INTRODUCTION

In Chapter 2 we introduced some sources of bias and variation that can occur during the conduct of clinical trials. The control of bias and variability is extremely important to ensure the integrity of clinical trials. In comparative clinical trials, randomization is usually used to control conscious or unconscious bias in the allocation of patients to treatment groups. The purpose of randomization is not only to generate comparable groups of patients who have similar characteristics but also to enable valid statistical tests for clinical evaluation of the study medicine.

The concept of randomization was first introduced in clinical research in the early 1930s for a study of sanocrysin in the treatment of patients with pulmonary tuberculosis (Amberson et al., 1931). However, the principle of randomization was not implemented in clinical trials until mid-1940s by the British Medical Research Council under the influence of Sir Austin Bradford Hill (1948). Since then there have been tremendous debates over the use of randomization in clinical research (e.g., see Feinstein, 1977, 1989). The primary concern is that it is not ethical for the patient not to know which treatment he or she receives, especially when one of the treatments is a placebo. However, it was not realized that before a clinical trial is conducted, no one can be 100% sure that the active treatment is indeed effective and safe for the indicated disease compared to the placebo. For many drug products it is not uncommon that the active treatment has inferior efficacy and safety than the placebo. One typical example would be the CAST study discussed in previous chapters. For another example, if randomization is not employed for comparing a surgical procedure with chemotherapy in treatment of patients with a certain cancer, then the so-called operable patients with good prognoses will more likely be assigned to surgery, while the chemotherapy will be given, as is usual, to the inoperable patients with poor prognoses.

The surgical treatment would have yielded the positive results even though the surgery was not performed at all.

The use of randomization can avoid subjective assignment of treatments to patients who participate in clinical trials. Its advantage can be best illustrated by clinical studies concerning the treatment of gastric freezing for patients with peptic ulcer conducted in the 1960s (Miao, 1977; Sackett, 1989). In these studies, the treatment of gastric freezing was applied to tens of thousands of patients with peptic ulcer in a nonrandom fashion. These studies showed that the gastric freezing might be a promising therapy for the disease. However, it only took one randomized trial with 160 patients, half to the real or sham freezing, to conclusively demonstrate that the treatment of gastric freezing is in fact ineffective for the treatment of peptic ulcer. Therefore, Section 314.166 of the CFR requires that the method for patient treatment assignment should be described in some detail in the study protocol and report. It is recommended that for a concurrent controlled study, treatment assignment of patients be done by randomization. It should be noted that randomization in clinical trials consists of (1) random selection of a representative sample from a targeted patient population and (2) random assignment of patients in order to study the medicines.

To remove the potential bias that might occur when there are inequalities between treatment groups (e.g., demographic details or prognostic variables) allocated to different treatment groups, the use of randomization with blocking and/or stratification, if necessary, is helpful. Lachin (1988a, 1988b) provides a comprehensive summary of the various randomization models. The concept behind these randomization models allows useful randomization methods to be employed such as the complete randomization, the permuted-block randomization, and the adaptive randomization. Randomization plays an important role for the generalization of the observed clinical trials. Therefore it is recommended that a set of standard operating procedures (SOP) for the implementation of randomization be developed when conducting clinical trials. In many clinical trials bias often occurs due to preconceived ideas or perceptions acquired during the study by (1) the investigator and supporting staff who might influence reporting response to therapy or adverse events and (2) the patient who might influence compliance, cooperation, or provision of information.

In clinical trials, in addition to randomization, the technique of blinding is usually employed to avoid the risk of personal bias in comparing treatments. Basically there are several different types of blinding commonly used in clinical trials. These blindings include open label (or unblinding), single blinding, double blinding, and triple blinding. An open label study indicates that both the patient and the investigator know to which treatment group the patient is assigned, while a single blinding is referred to as that when the investigator knows but the patient does not. For a double blinding, neither the investigator nor the patient knows to which treatment group the patient is assigned. A triple blinding is an extension of the double blinding in which those monitoring outcome are unaware of treatment assignment. In practice, randomization and blinding are important to the success of clinical trials. Randomization and blinding can not only help to avoid bias but also to control variability, and consequently to achieve the desired accuracy and reliability of clinical trials.

The remainder of this chapter is organized as follows. In the next section, we introduce various randomization models. Section 4.3 covers the different randomization methods. In Section 4.4 we provide a commonly employed approach for the implementation of randomization in the pharmaceutical industry. The issue regarding the generalization of controlled randomized trials is discussed in Section 4.5. The concept for the use of blinding in clinical trials is addressed in Section 4.6. A brief discussion is given in Section 4.7.

4.2 RANDOMIZATION MODELS

As was indicated in the preceding chapter, randomization ensures that patients selected from the target patient population constitute a representative sample of the target patient population. Therefore, statistical inference can be drawn based on some probability distribution assumption of the target patient population. The probability distribution assumption depends on the method of randomization under a randomization (population) model. As a result a study without randomization will result in the violation of the probability distribution assumption, and consequently no accurate and reliable statistical inference on the study medicine can be drawn.

Lachin (1988a) provides a comprehensive summary of the randomization basis for statistical tests under various models. His observations are discussed below.

Population Model

Cochran (1977) points out that the validity of statistical inference by which clinicians can draw conclusions for the patient population is based on the selection of a representative sample drawn from the patient population by some random procedure. This concept is called the *population model* (Lehmann, 1975; Lachin, 1988a). Suppose that for a certain disease, a clinical trial is planned to investigate the efficacy and safety of a newly developed therapeutic agent compared to an inert placebo. Under the population model we can draw two samples independently with equal chance at random from the (infinitely large) patient population. One sample consists of n_T patients, and the other sample consists of n_P patients. We denote these two samples by sample T and sample P , respectively. The n_T patients in sample T will receive the newly developed agent, while the inert placebo is given to the n_P patients in sample P . If the patient population is homogeneous with respect to the inclusion and exclusion criteria specified in the protocol, we do not expect that the responses of clinical endpoints for a particular patient will have anything to do with those of other patients. In other words, they are statistically independent of one another. For a homogeneous population, a common (population) distribution can be used to describe the characteristics of the clinical responses. That is, they are assumed to have identical distribution. Hence the clinical responses of the n_I patients ($I = T, P$) are said to have an independent and identical distribution (i.i.d.). Therefore, optimal statistical inference can be precisely obtained. For example, with respect to hypotheses (2.6.1) regarding the detection of the difference between the new agent and the placebo, the common two-sample t -test is the optimal testing procedure (Armitage and Berry, 1987).

As mentioned above, randomization in clinical trials involves random selection of the patients from the population and random assignment of patients to the treatments. Under the assumption of a homogeneous population, the clinical responses of all patients in the trial, regardless of sample T or sample P , are independent and have the same distribution. Lachin (1988a) points out that the significance level (i.e., the probability of type I error) and the power (i.e., the probability of correctly rejecting a false null hypothesis) will not be affected by random assignment of patients to the treatments as long as the patients in the trial represent a random sample from the homogeneous population. Furthermore suppose that we split a random sample into two subsamples; the statistical inferential procedures are still valid even if the one-half of the patients are assigned to the test drug and the other half to the placebo.

Invoked Population Model

In clinical trials, we usually select investigators first and then select patients at each selected investigator's site. At each selected study site, the investigator will usually enroll qualified patients sequentially. A qualified patient is referred to as a patient who meets the inclusion and exclusion criteria and has signed the informed consent form. As a result, neither the selection of investigators (or study centers) nor the recruitment of patient is random. However, patients who enter a trial are assigned to treatment groups at random. In practice, the collected clinical data are usually analyzed as if they were obtained under the assumption that the sample is randomly selected from a homogeneous patient population. Lachin (1988a) refers to this process as the *invoked population model* because the population model is invoked as the basis for statistical analysis as if a formal sampling procedure were actually performed. In current practice, the invoked population model is commonly employed for data analysis for most clinical trials. It, however, should be noted that the invoked population model is based on the assumption that it is inherently untestable.

Note that one of the underlying assumptions for both the population model and the invoked population model is that the patient population is homogeneous. This assumption, however, is not valid in most clinical trials. In practice, we can employ the technique of stratified sampling to select samples according to some prespecified covariates to describe the differences in patient characteristics. The idea of stratification is to have homogeneous subpopulations with respect to the prespecified covariates (or patient characteristics). In many clinical trials, it is almost impossible to use a few covariates to describe the differences among heterogeneous subpopulations due to the complexity of patient characteristics and disease conditions. In addition, patients who are enrolled at different times may not have similar relevant demographic and baseline characteristics. In other words, the patient population is time-heterogeneous population in which the patient's characteristics are a function of the time when they enter the trial. The impact of the heterogeneity due to the recruitment time on the results of a clinical trial is well documented in the literature. For example, Byar et al. (1976) indicate that in a study conducted by the Veterans Administration Cooperative Urological Research Group in 1967, the survival rate of the patients who entered earlier in the study was worse than of those who enrolled later in the study. Therefore it is not uncommon for patient characteristics to change over time even if the population is homogeneous at one time point. The above discussion indicates that the assumption of the population model or the invoked population model may not be valid.

Randomization Model

As discussed above, for current practice, although the study site selection and patient selection are not random, the assignment of treatments to patients is usually performed based on some random mechanism. Thus, treatment comparisons can be made based on the so-called randomization or permutation tests introduced in the mid-1930s (Fisher, 1935). To illustrate the concept of permutation tests, we consider the following hypothetical data set concerning endpoint changes from baselines in peak urinary flow rate (mL/s) after three months of treatment for patients with benign prostate hyperplasia:

Test drug : 2.6, 0.97, 1.68;

Placebo : 1.2, -0.43.

Table 4.2.1 All Possible Ranks for the Two Patients in the Placebo Group Based on Conditional Permutation

Possible Ranks	Sum of Ranks
1, 2	3
1, 3	4
1, 4	5
1, 5	6
2, 3	5
2, 4	6
2, 5	7
3, 4	7
3, 5	8
4, 5	9

An interesting question is how to determine whether there is a significant difference in endpoint change from baseline in peak urinary flow rate between the test drug and the placebo based on the above hypothetical data. Under the null hypothesis of no difference described in (2.6.1), all possible permutations according to the endpoint changes from baselines in peak urinary flow rate are equally likely (from the smallest to the largest based in the ranking). If all possible pairs of ranks for the two patients receiving placebo are all equally likely, then the sum of the ranks for the two patients in the placebo group are also equiprobable. The possible ranks and sum of the ranks for the two patients receiving placebo are given in Table 4.2.1. Since the chance is equal for all possible permutations of the ranks for the two patients in the placebo group, the probability distribution for the sum of the ranks can be obtained as given in Table 4.2.2. Since the ranks of the observed endpoint change from baseline in peak urinary flow rate for the two patients in the placebo group are 3 (for 1.2) and 1 (for -0.73), respectively, the rank sum for the placebo group is 4. As can be seen from Table 4.2.2, the p -value (i.e., the probability that the observed rank sum is due to chance or the sum of the ranks from placebo group can be at least as extreme as the observed 4) is 0.2 for a one-sided test and 0.4 for a two-sided test. Note that the possible values and the distribution of the rank sums will be the same no matter what actual observed endpoint change from baseline in peak urinary flow rate for the two patients in the placebo group are as long as the two patients are assigned to the placebo

Table 4.2.2 Probability Distribution of the Sum of Ranks Based on Conditional Permutation

Sum of Ranks	Probability
3	0.1
4	0.1
5	0.2
6	0.2
7	0.2
8	0.1
9	0.1

Table 4.2.3 All Possible Unconditional Permutation for Five Subjects

Number of Subjects		
Placebo	Test	Possible Permutation
0	5	1
1	4	5
2	3	10
3	2	10
4	1	5
5	0	1
Total		$32 = 2^5$

group at random. The above test is known as the Wilcoxon rank sum test (Wilcoxon, 1945). The Wilcoxon rank sum test is one of the conditional permutation tests in which permutation is performed to the confinement of random assignment of two out of five patients with prostate hyperplasia to the placebo group. As a result, there are a total of 10 possible permutations of ranks for the placebo group. If we can randomly assign any five enrolled patients from 0 to 5 to receive placebo treatment, then there will be a total of 32 subsets as given in Table 4.2.3. The permutation tests over all possible subsets are called the unconditional permutation tests. The *p*-values for the unconditional permutation tests can be similarly computed.

The above discussion indicates that the calculation of the *p*-value for the Wilcoxon rank sum test does not assume any probability distribution for the endpoint change from baseline in peak urinary flow rate from baseline. As a matter of fact, any statistical test based on the permutation principle is assumption free. In addition, as indicated by Lachin (1988a), the family of the linear rank tests is the most general family of permutation tests (Lehmann, 1975; Randles and Wolfe, 1979). For example, the well-known Pearson chi-square statistic for comparison of two proportions is equal to $N(N-1)$ times the chi-square statistic derived by permutation, where N is the total number of patients (Koch and Edwards, 1988). Other tests for continuous or quantitative clinical endpoints include the Wilcoxon rank sum test for two independent samples and Kruskal-Wallis test for multiple independent samples. For censored data the logrank test (Miller, 1981) and Peto-Peto-Prentice-Wilcoxon test (Kalbfleisch and Prentice, 1980) are useful. These tests are widely applied statistical procedures in clinical trials. Note that since the statistical procedures based on the concept of permutation require the enumeration of all possible permutations, it is feasible only for small samples. As the sample size increases, however, the sampling distribution of test statistics derived under permutation will approach to some known continuous distribution such as a normal distribution. In addition, as shown by Lachin (1988b), the probability distributions for the family of linear rank statistics for large samples are equivalent to those of the tests obtained under the assumption of the population model. As a result, if patients are randomly assigned to the treatments, statistical tests for evaluation of treatments should be based on permutation tests because the exact *p*-value can be easily calculated for small samples. For large samples, the data can be analyzed by the permutation methods derived under the population model as if the patients were randomly selected from a homogeneous population.

Stratification

In most clinical trials, the ultimate goal is not only to provide statistical inference on the effectiveness and safety of a test drug, compared to a control or placebo based on clinical data collected from the trials, but also to apply the results to the targeted patient population. In practice, there are many covariates such as age, gender, race, geographical locations, underlying disease severity, and others that may have an impact on the statistical inference drawn. The accuracy and reliability of the estimation of primary clinical endpoints for evaluation of the treatment effect can be affected by the heterogeneity caused by these covariates. To overcome and control such heterogeneity, a stratified randomization is found helpful. The use of stratification in clinical trials is motivated originally by the concept of blocking in agricultural experiments in the mid-1930s (Fisher, 1935). The idea is quite simple and straightforward. If a covariate is known to be the cause of heterogeneity, then the patients are stratified or blocked into several homogeneous groups (or strata) with respect to the covariate. Randomization of patients to the treatment is then performed independently within the strata. This type of randomization with strata is called *stratified randomization*. For example, the National Institute of Neurological Disorders and Stroke rt-PA stroke study group (1995) suspected that the time from the onset of stroke to the beginning of treatment of rt-PA may have a significant impact on neurologic improvement as assessed by the National Institute of Health Stroke Scale (NIHSS). As a result the study considered two strata of patients based on the time (in minutes) from the onset to the start of the treatment, namely 0 to 90 minutes and 91 to 180 minutes. For multicenter trials, stratified randomization with respect to geographical location is necessary because differences in study centers usually account for the major source of variation for many primary clinical endpoints. The idea of stratification is to keep the variability of patients within strata as small as possible and the between-strata variability as large as possible so that the inference for the treatment effect possesses the optimal precision. Another reason for the use of stratification in clinical trials is to prevent imbalance with respect to important covariates. For example, with an unstratified randomization, more males may be enrolled into the test drug group, while the placebo group may enroll more females. Hence the distribution of treatments with respect to gender is not balanced. Despite the advantages of stratified randomization, it should be noted that the stratification will eventually become more complicated and difficult to implement due to administrative complexity, increasing time and expense, and other logistic issues.

The extreme case of stratification is the technique of matching which is often employed in the case-control studies. For example, for a clinical trial comparing a test drug with a placebo, patients are to be matched in pairs with respect to some predetermined covariates such as demographic, baseline characteristics, and severity of disease. Within each pair, one patient is assigned to receive the test drug and the other patient receives the placebo. Assignment of the matched patients to treatments might not be random. Wooding (1994) points out that matching is often used as a substitution for randomization by investigators who mistrust and do not like the concept of randomization. In case-control trials, although there may be a large number of covariates, they may or may not have an impact on clinical outcomes. In practice, it is almost impossible to consider all possible covariates in a clinical trial. However, if an important covariate is missed during the process of matching, the cause of bias cannot be identified, and consequently it cannot be assessed if it truly exists. Another problem of matching in case-control trials is that the number of patients increases rapidly as the number of covariates to be considered for matching becomes large.

Accordingly, the task of finding matching pairs becomes formidable (e.g., see Wooding, 1994). The primary purpose of matching is to eliminate variations of clinical endpoints caused by the differences among patients due to biological variations, as discussed in Chapter 2. As a result the method of matching attempts to consider the individual patient as a stratum and to randomize the sequence of treatment of the test drug and placebo within each stratum. Since each patient receives both treatments, the variation between patients is eliminated from the comparison between the test drug and the placebo. This type of design is called a crossover design, and it will be discussed in detail in the next chapter.

In summary, as indicated by Lachin (1988a), the chance of covariate imbalance decreases as the sample size increases. The covariate imbalance has little impact on large samples. In addition the difference in statistical power between unstratified and stratified randomization is negligible (McHugh and Matts, 1983). Furthermore a post hoc stratified analysis can always be employed to adjust bias caused by the imbalance of baseline covariates. To control covariate imbalance, Peto et al. (1976) indicate that if, during analysis, initial diagnosis (i.e., covariate) is allowed for (i.e., stratified analysis) as the different treatments are being compared, there is hardly ever need for stratification at entry in large trials. Therefore it is recommended that stratified randomization for a clinical trial be performed only with respect to those covariates that are absolutely necessary for the integrity of the study. In addition, ICH E9 guidance on *Statistical Principles for Clinical Trials* also suggests that although stratified randomization by important prognostic variables may be valuable to keep balanced allocation within strata and has greater potential benefit in small trials, the use of more than two or three stratified variables is rarely necessary and is logically troublesome (ICH E9, 1998).

4.3 RANDOMIZATION METHODS

In the early 1970s, before the concept of randomization was widely accepted as an effective tool to prevent the subjective selection bias in the assignment of patients to the treatments, some systematic methods for assignment of patients to treatments under study were commonly used. These systematic methods are summarized in Table 4.3.1. As can be seen from Table 4.3.1, all of these methods are deterministic. The assignment of patients to treatments can be predicted without error. Since the investigators or patients may be aware of which treatment the patients receive, subjective bias can consciously or unconsciously occur in both the assignment of patients to treatments and the evaluation of clinical outcomes for the treatment under investigation. To prevent such bias, in this section several useful randomization methods are introduced.

Although controlled randomized trials are viewed as the state-of-the-art technology for clinical evaluation of therapeutic interventions, some investigators still try to beat the randomization by guessing the treatments to which the patients are assigned (Karlowski et al.,

Table 4.3.1 Unacceptable Methods of Assignment of Patients to Treatment

-
1. Assignment of patients to treatment according to the order of enrollment
(every other patient is assigned to one group)
 2. Assignment of patients to treatment according to patient's initial
 3. Assignment of patients of treatment according to patient's birthday
 4. Assignment of patients according to the dates of enrollment
-

Table 4.3.2 Blackwell-Hodges Diagram for Selection Bias

Investigator's Guess	Random Assignment of Equal Probability	
	Test Drug	Placebo
Test drug	a	$n/2 - b$
Placebo	$n/2 - a$	b
	$n/2$	$n/2$

Source: Blackwell and Hodges (1957).

1975; Byington et al., 1985; Deyo et al., 1990). Hence, subjective judgment for evaluation of patients' clinical outcomes always introduces potential selection bias by investigators who are aware of the treatment assignment of patients. A simple model suggested by Blackwell and Hodges (1957) and Lachin (1988a) can be adapted to assess this potential selection bias due to a wrong guess of treatment assignments by investigators. The Blackwell-Hodges diagram for selection bias is given in Table 4.3.2. This diagram is constructed under the assumption that each patient has an equal chance (50%) of being assigned to either the test drug or the placebo. Therefore, if there are a total of n patients enrolled into the study, the expected sample size for both the test drug and the placebo is equal to $n/2$. Then, the total potential selection bias for evaluation of the treatment effect introduced by the investigator is represented as the (expected) difference between the observed sample means according to the treatment assignments guessed by the investigator. This expected difference can then be shown as the product of the investigator's bias in favor of the test drug times the expected bias factor which is the difference between the expected number of correct guesses and the number expected by chance. The expected bias factor is equal to one-half times the number of correct guesses minus the number of misses. Suppose that the study is double blinded and that the investigators have no other way to predict the treatment assignments but to use laboratory evaluations or some particular adverse events caused by the test drug. Then, the probability of correctly guessing the treatment assignments is 50% for each treatment. Consequently, under this situation, the expected number of correct guesses will be the same as the number expected by chance, which is $n/2$. Hence the expected bias factor is zero. Therefore, even though the investigator might have positive bias in evaluation of patients whom he or she believe are receiving the test drug, the potential selection bias will vanish in the evaluation of the treatment effect due to the fact that the expected bias factor is zero.

Note that in addition to selection bias, an accidental bias can also occur when comparing treatments in the presence of covariate imbalances. Efron (1971) considers the effects of various randomization methods on bias for estimation of the treatment effect in a regression model assuming that important covariates are not accounted for. Gail et al. (1984) and Lachin (1988a) reported that these randomization methods will generally produce consistent estimates of the treatment effect in linear models. However, in some nonlinear models estimates of the treatment effect are biased no matter how large the sample size is (i.e., asymptotically biased) under these randomization methods. Note that for linear models, these randomization methods are equivalent in the sense that they produce estimates of treatment effect that are free of accidental bias. However, for small or finite samples, the

Table 4.3.3 Example of Complete Randomization for Four Centers

Random codes for drug XXX, protocol XXX-014
Double-blind, randomized, placebo-control, two parallel groups

A. Hope, MD		J. Smith, MD	
Subject Number	Treatment Assignment	Subject Number	Treatment Assignment
1403001	Active drug	1401001	Placebo
1403002	Active drug	1401002	Active drug
1403003	Placebo	1401003	Active drug
1403004	Active drug	1401004	Active drug
1403005	Placebo	1401005	Placebo
1403006	Active drug	1401006	Placebo
1403007	Active drug	1401007	Active drug
1403008	Placebo	1401008	Active drug
1403009	Active drug	1401009	Placebo
1403010	Placebo	1401010	Placebo
1403011	Placebo	1401011	Placebo
1403012	Placebo	1401012	Active drug
1403013	Active drug	1401013	Placebo
1403014	Placebo	1401014	Active drug
1403015	Placebo	1401015	Placebo
1403016	Active drug	1401016	Active drug
1403017	Placebo	1401017	Active drug
1403018	Placebo	1401018	Placebo
1403019	Active drug	1401019	Active drug
1403020	Placebo	1401020	Active drug
1403021	Active drug	1401021	Placebo
1403022	Active drug	1401022	Active drug
1403023	Placebo	1401023	Active drug
1403024	Active drug	1401024	Active drug

variance of the bias varies from randomization method to randomization method. As a result the chance of accidental bias and magnitude of accidental bias vary with respect to randomization methods.

In general, randomization methods can be classified into three types according to the restriction of the randomization and the change in probability for randomization with respect to the previous treatment assignments. These types of randomization methods are the complete randomization, the permuted-block randomization, and the adaptive randomization. Randomization can be performed either by random selection or by random allocation for methods of complete and permuted-block randomization. Basically the adaptive randomization consists of treatment and covariate and response adaptive randomizations. In what follows we will describe these randomization methods and compare their relative merits and limitations whenever possible.

The randomization list of a clinical trial documents the random assignment of treatments to subjects. As presented in Tables 4.3.3 to 4.3.4, it is a sequential list of treatments or treatment sequences in a crossover trial, or corresponding codes by subject numbers. As

Table 4.3.3 Example of Complete Randomization for Four Centers (*Continued*)

Random codes for drug XXX, protocol XXX-014
Double-blind, randomized, placebo-control, two parallel groups

M. Dole, MD		C. Price, MD	
Subject Number	Treatment Assignment	Subject Number	Treatment Assignment
1402001	Placebo	1404001	Placebo
1402002	Active drug	1404002	Active drug
1402003	Placebo	1404003	Active drug
1402004	Placebo	1404004	Placebo
1402005	Active drug	1404005	Active drug
1402006	Active drug	1404006	Placebo
1402007	Active drug	1404007	Placebo
1402008	Active drug	1404008	Active drug
1402009	Active drug	1404009	Placebo
1402010	Active drug	1404010	Active drug
1402011	Active drug	1404011	Active drug
1402012	Placebo	1404012	Active drug
1402013	Placebo	1404013	Placebo
1402014	Placebo	1404014	Placebo
1402015	Active drug	1404015	Placebo
1402016	Active drug	1404016	Placebo
1402017	Placebo	1404017	Active drug
1402018	Active drug	1404018	Placebo
1402019	Active drug	1404019	Placebo
1402020	Active drug	1404020	Active drug
1402021	Placebo	1404021	Active drug
1402022	Placebo	1404022	Placebo
1402023	Active drug	1404023	Placebo
1402024	Placebo	1404024	Placebo

different trials might have different study designs, different objectives, or different prognostic factors to consider, different procedures for generating randomization codes might be necessary. In addition, the randomization codes should be reproducible. Once the trial starts, subjects who meet the inclusion and exclusion criteria should receive their corresponding random treatment assignment according to the randomization codes sequentially. In other words, the next subject to be randomized into a clinical trial should always receive the treatment to the next free number in the appropriate randomization list. However, too much detailed information of randomization will facilitate predictability and should not be included in the protocol. Furthermore, the randomization list should be filed securely in a manner that blindness is adequately maintained throughout the study.

Complete Randomization

Simple randomization is referred to as the procedure in which no restrictions are enforced on the nature of randomization sequence except for the number of patients required for achieving

Table 4.3.4 Example of Complete Randomization for Four Centers

Random codes for drug XXX, protocol XXX-014
Double-blind, randomized, placebo-control, three parallel groups

A. Hope, MD		J. Smith, MD	
Subject Number	Treatment Assignment	Subject Number	Treatment Assignment
1403001	Placebo	1401001	Placebo
1403002	100 mg	1401002	200 mg
1403003	100 mg	1401003	100 mg
1403004	Placebo	1401004	100 mg
1403005	200 mg	1401005	100 mg
1403006	200 mg	1401006	Placebo
1403007	100 mg	1401007	200 mg
1403008	200 mg	1401008	Placebo
1403009	100 mg	1401009	200 mg
1403010	100 mg	1401010	200 mg
1403011	100 mg	1401011	Placebo
1403012	200 mg	1401012	Placebo
1403013	Placebo	1401013	200 mg
1403014	100 mg	1401014	Placebo
1403015	Placebo	1401015	200 mg
1403016	200 mg	1401016	200 mg
1403017	Placebo	1401017	200 mg
1403018	100 mg	1401018	100 mg
1403019	200 mg	1401019	200 mg
1403020	100 mg	1401020	Placebo
1403021	Placebo	1401021	100 mg
1403022	Placebo	1401022	200 mg
1403023	Placebo	1401023	100 mg
1403024	Placebo	1401024	100 mg

the desired statistical power and the ratio of patient allocation between treatments. For a clinical trial with N patients comparing a test drug and a placebo, the method of simple randomization is called a completely binomial design (Blackwell and Hodges, 1957) or a simply complete randomization (Lachin, 1988b) if it has the following properties:

1. The chance that a patient receives either the test drug or the placebo is 50%.
2. Randomization of assignments is performed independently for each of the N patients.

The randomization codes based on the method of complete randomization can be generated either by the table of random numbers (Pocock, 1983) or by some statistical computing software such as SAS® (Statistical Analysis System, 1995). However, it should be realized that a computer cannot generate *true* random numbers but *pseudorandom* numbers because only a fixed number of different long series of almost unpredictable permuted numbers are generated. Therefore, Lehmann (1975) recommends that a run test be performed to verify the randomness of the generated randomization codes. In practice, however, randomization

Table 4.3.4 Example of Complete Randomization for Four Centers (*Continued*)

Random codes for drug XXX, protocol XXX-014

Double-blind, randomized, placebo-control, three parallel groups

M. Dole, MD		C. Price, MD	
Subject Number	Treatment Assignment	Subject Number	Treatment Assignment
1402001	Placebo	1404001	100 mg
1402002	100 mg	1404002	200 mg
1402003	100 mg	1404003	Placebo
1402004	Placebo	1404004	Placebo
1402005	100 mg	1404005	200 mg
1402006	100 mg	1404006	200 mg
1402007	200 mg	1404007	200 mg
1402008	100 mg	1404008	100 mg
1402009	100 mg	1404009	Placebo
1402010	200 mg	1404010	200 mg
1402011	Placebo	1404011	100 mg
1402012	100 mg	1404012	Placebo
1402013	200 mg	1404013	200 mg
1402014	Placebo	1404014	Placebo
1402015	200 mg	1404015	Placebo
1402016	Placebo	1404016	100 mg
1402017	Placebo	1404017	200 mg
1402018	Placebo	1404018	200 mg
1402019	100 mg	1404019	100 mg
1402020	Placebo	1404020	Placebo
1402021	100 mg	1404021	100 mg
1402022	200 mg	1404022	200 mg
1402023	200 mg	1404023	200 mg
1402024	Placebo	1404024	200 mg

codes are preferably generated by a computer due to its speed and convenience in the maintenance of generated randomization codes. For example, the SAS® function RANBIN can be used to generate randomization codes for clinical trials with two treatment groups. For another example, suppose that a clinical trial is planned in four study centers to investigate the effectiveness and safety of a test drug as compared to an inert placebo. Ninety-six patients are intended for the study. Suppose that it is desirable to allocate patients equally in each treatment group by study center. We will consider the randomization codes given in Table 4.3.3 as generated based on complete randomization by study center. Suppose that there are three treatment groups (e.g., placebo, 100 mg, and 200 mg of the test drug), and the randomization codes can be similarly generated (see Table 4.3.4). Note that the SAS programs used for generation of the randomization codes given in Tables 4.3.3 and 4.3.4 are provided in Appendices B.1 and B.2, respectively.

In clinical trials, in the interest of balance, the assignment of an equal number of patients in treatment groups is usually considered. In practice, however, it is possible that a trial will end up with an unequal number of patients in each treatment group. Table 4.3.5 provides a distribution of the number of patients by treatment and study center for the

Table 4.3.5 Sample Size by Treatment and Center for Random Codes in Tables 4.3.3 and 4.3.4

Study Center	Active Drug	Placebo	Total	
<i>Table 4.3.3</i>				
J. Smith, M.D.	14	10	24	
M. Dole, M.D.	14	10	24	
A. Hope, M.D.	12	12	24	
C. Price, M.D.	10	14	24	
Total	50	46	96	
Study Center	100 mg	200 mg	Placebo	
<i>Table 4.3.4</i>				
J. Smith, M.D.	7	10	7	24
M. Dole, M.D.	9	6	9	24
A. Hope, M.D.	9	6	9	24
C. Price, M.D.	6	11	7	24
Total	31	33	32	96

randomization codes given in Tables 4.3.3 and 4.3.4. It can be seen that although the probability for random assignments is 1/2 for two treatments and 1/3 for three groups, the final sample sizes based on complete randomization are not equal for the treatment groups. In general, the treatment imbalance within each study center is more severe than that for the overall clinical trial. Lachin (1988b) provides an approximate formula for calculation of the chance of treatment imbalance for complete randomization. Based on his formula, Figure 4.3.1 plots the chance of treatment imbalance as a function of sample size for different fractions of the total sample size for the larger treatment group. It can be seen from Figure 4.3.1 that the minimum sample sizes required for a probability of less than 5% for the treatment imbalance are 386, 96, 44, and 24 when the fractions of the total sample size for the larger treatment (imbalance proportion) are 0.55, 0.60, 0.65, and 0.70, respectively. On the other hand, Figure 4.3.2 gives a graphical presentation of the fractions of the total sample size for a larger group such that would occur with the probabilities 0.005, 0.01, and 0.05 as a function of sample sizes. Figure 4.3.2 clearly shows that the fraction of the total sample size for the large group with a fixed probability of treatment imbalance is a decreasing function of sample size. Both Figures 4.3.1 and 4.3.2 demonstrate that a severe treatment imbalance based on complete randomization is unlikely when the sample size is large. For the usual statistical tests for quantitative clinical measures that can adequately be described by the normal probability model, the smallest variance for the estimate of the treatment effect can be obtained when an equal number of patients are enrolled in each treatment group. Consequently, the maximum statistical power for detection of the treatment difference is achieved. In order to examine the impact on the power caused by treatment imbalance due to the complete randomization, Figure 4.3.3 provides a graph of the power for detection of a fixed treatment difference as a function of the fraction of a fixed total sample size for the larger treatment group. As can be seen from Figure 4.3.3, when the fraction of the larger group is at most 0.7, the power of the test for detection of treatment effect is hardly affected at all. In summary, although complete randomization

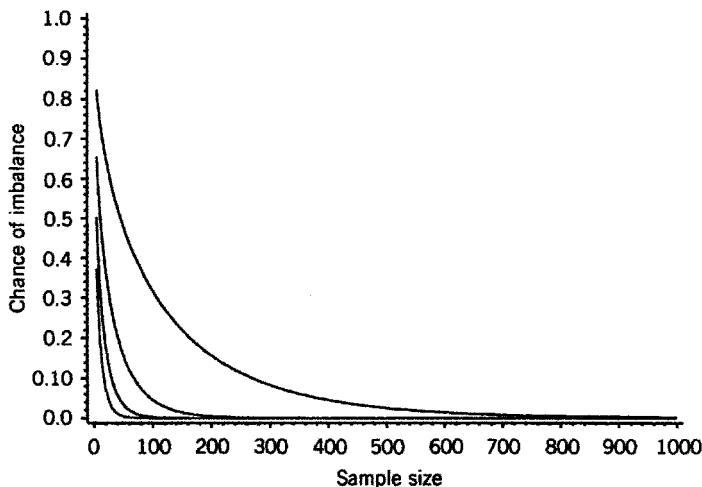


Figure 4.3.1 Chance of treatment imbalance for complete randomization as a function of sample size. Sample fraction: 0.55, 0.60, 0.65, and 0.70. (Source: Lachin, 1988b.)

will present a high chance of treatment imbalance, the chance of severe treatment imbalance is unlikely and moderate treatment imbalance has little impact on statistical power if the sample size of the trial exceeds 200. In practice, complete randomization is easy to implement. However, there is a high probability that it will produce unequal sample sizes among treatment groups when the total sample size is moderate (e.g., fewer than a few hundreds).

Another type of simple randomization that provides equal allocation of sample size is the random allocation. Random allocation is the simplest form of restricted randomization.

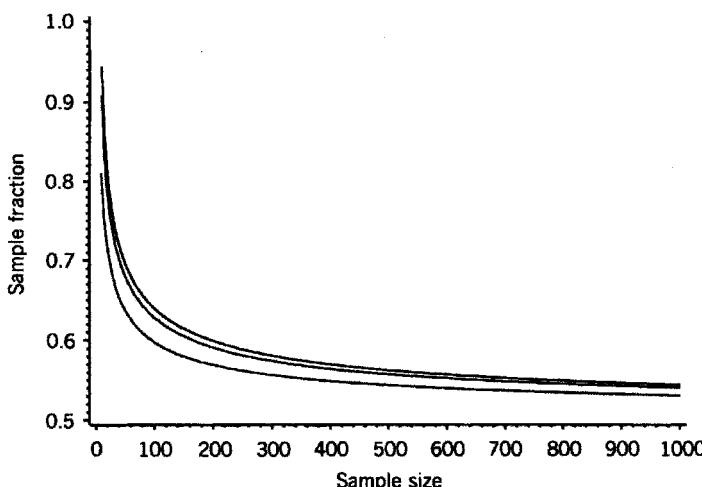


Figure 4.3.2 Sample fractions for complete randomization with chance of imbalance as a function of sample size. Chance of imbalance: 0.005, 0.01, and 0.05. (Source: Lachin, 1988b.)

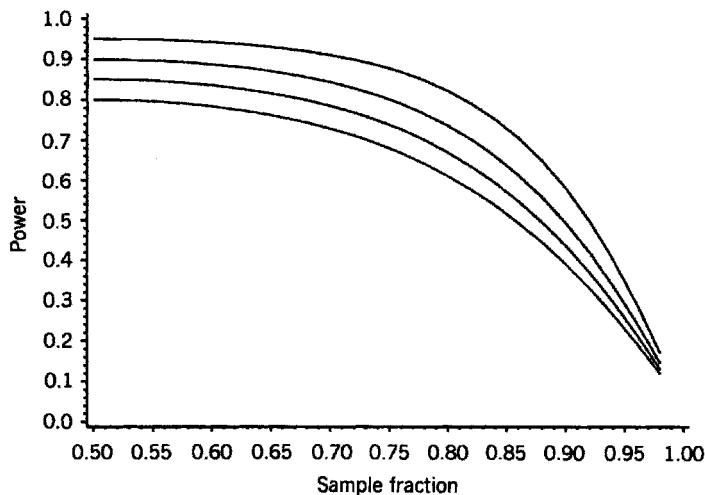


Figure 4.3.3 Power curves as a function of the sample fractions. Power is for a two-sided test at the 5% significance level. Power is 0.95, 0.90, 0.85, and 0.80 when the sample fraction is 0.5. (Source: Lachin, 1988a.)

The method of random allocation randomly selects the $N/2$ out of a total of N patients without replacement and assigns these $N/2$ patients to receive the test drug and the other half to receive the placebo. Since there are a total of $N!/[{(N/2)!]^2}$ possible ways for the selection of $N/2$ patients, it is equivalent to generating a random permutation of numbers from 1 to N and assigning the first half to the test drug. Hence, the SAS® procedure PLAN can be used to generate randomization codes for the method of random allocation. Table 4.3.6 provides a list of the randomization codes for a test drug and a placebo; the codes are generated based on a method of random allocation by the study center. It can be verified that for each of the four centers exactly 12 patients are assigned to each treatment group. Thus a total of 48 patients are assigned to either the test drug group or the placebo group. An SAS program for the method of random allocation for simple randomization is also provided in Appendix B.3. Although the marginal probability for assigning a patient to each of the two treatment groups is 1/2 for the method of random allocation, the conditional probability for assignment of a patient given that the assignment of the previous patient is not equal to 1/2 for the method of random allocation. This is because the random allocation is based on simple sampling without replacement.

Note that in a unblinded study, there is no potential selection bias for complete randomization, since the expected bias factor is always zero. As indicated in Lachin (1988b), the expected bias factor under the method of random allocation is an increasing function of sample size. As a result the selection bias for the method of random allocation can be very substantial as the sample size increases. Therefore it is extremely important to keep the study double blinded if the method of random allocation is employed. With respect to accidental bias caused by omitting some important covariates in estimating the treatment effect, both methods of complete randomization and random allocation are insensitive to covariate imbalance and hence are free of accidental bias when sample size is large, say over 100. For small samples the accidental bias may potentially exist for both methods. However, the accidental bias generated by the method of random allocation is larger than

Table 4.3.6 Example of Random Allocation for Four Centers

Random codes for drug XXX, protocol XXX-014

Double-blind, randomized, placebo-control, two parallel groups

A. Hope, MD			J. Smith, MD		
Subject Number	Random Permutation	Treatment Assignment	Subject Number	Random Permutation	Treatment Assignment
1403001	15	Placebo	1401001	19	Placebo
1403002	12	Active drug	1401002	9	Active drug
1403003	19	Placebo	1401003	22	Placebo
1403004	1	Active drug	1401004	17	Placebo
1403005	23	Placebo	1401005	16	Placebo
1403006	11	Active drug	1401006	5	Active drug
1403007	2	Active drug	1401007	8	Active drug
1403008	20	Placebo	1401008	21	Placebo
1403009	3	Active drug	1401009	13	Placebo
1403010	22	Placebo	1401010	12	Active drug
1403011	10	Active drug	1401011	24	Placebo
1403012	16	Placebo	1401012	6	Active drug
1403013	4	Active drug	1401013	4	Active drug
1403014	6	Active drug	1401014	14	Placebo
1403015	7	Active drug	1401015	1	Active drug
1403016	13	Placebo	1401016	15	Placebo
1403017	24	Placebo	1401017	10	Active drug
1403018	9	Active drug	1401018	3	Active drug
1403019	17	Placebo	1401019	7	Active drug
1403020	21	Placebo	1401020	23	Placebo
1403021	18	Placebo	1401021	2	Active drug
1403022	8	Active drug	1401022	20	Placebo
1403023	5	Active drug	1401023	18	Placebo
1403024	14	Placebo	1401024	11	Active drug

that of complete randomization. As a matter of fact the accidental bias under complete randomization is the smallest among all randomization methods discussed in this section.

Permuted-Block Randomization

One of the major disadvantages of simple randomization is that treatment imbalance can occur periodically. For example, in Table 4.3.3, the randomization codes for investigator M. Dole, M.D., were generated under complete randomization with a run of seven consecutive patients (from subject 1402005 to subject 1402011) assigned to receive the test drug. Two observations should be noted. First, the number of patients in the treatment groups is not balanced, as was previously shown in Table 4.3.5. This may be in part explained by the fact that the randomization is performed within each center and the number of patients to be enrolled at each center is usually fewer than 50. Second, most clinical trials recruit patients sequentially. If the demographic factors or baseline characteristics change over time, then it is quite possible to have a serious covariate imbalance between treatment groups within each study center and for the entire study as well. This covariate imbalance

Table 4.3.6 Example of Random Allocation for Four Centers (*Continued*)

Random codes for drug XXX, protocol XXX-014
Double-blind, randomized, placebo-control, two parallel groups

M. Dole, MD			C. Price, MD		
Subject Number	Random Permutation	Treatment Assignment	Subject Number	Random Permutation	Treatment Assignment
1402001	23	Placebo	1404001	4	Active drug
1402002	21	Placebo	1404002	17	Placebo
1402003	13	Placebo	1404003	15	Placebo
1402004	17	Placebo	1404004	20	Placebo
1402005	7	Active drug	1404005	5	Active drug
1402006	10	Active drug	1404006	3	Active drug
1402007	18	Placebo	1404007	14	Placebo
1402008	20	Placebo	1404008	10	Active drug
1402009	1	Active drug	1404009	11	Active drug
1402010	14	Placebo	1404010	19	Placebo
1402011	19	Placebo	1404011	21	Placebo
1402012	3	Active drug	1404012	2	Active drug
1402013	22	Placebo	1404013	22	Placebo
1402014	9	Active drug	1404014	23	Placebo
1402015	24	Placebo	1404015	16	Placebo
1402016	5	Active drug	1404016	7	Active drug
1402017	16	Placebo	1404017	9	Active drug
1402018	8	Active drug	1404018	8	Active drug
1402019	4	Active drug	1404019	1	Active drug
1402020	15	Placebo	1404020	24	Placebo
1402021	11	Active drug	1404021	18	Placebo
1402022	6	Active drug	1404022	6	Active drug
1402023	2	Active drug	1404023	12	Active drug
1402024	12	Active drug	1404024	13	Placebo

can be potentially disastrous. For example, suppose that a clinical trial is conducted in two study centers to evaluate the effectiveness and safety of a test drug as compared to a placebo. A complete randomization is used for the generation of randomization codes. Suppose that the randomization codes for one of the two centers contain a long run of consecutive patients who are assigned to the test drug group. Also suppose that one of the important baseline covariates is not balanced between the two treatments within the center. In this case it is extremely difficult to explain a possible difference in treatment effect between the two centers because the center effect is confounded with the effect due to this covariate. One resolution to this major disadvantage of simple randomization is periodically to enforce a balance in the number of patients assigned to each treatment. In other words, we first divide the whole series of patients who are to enroll in the trial into several blocks with equal or unequal lengths. We then randomize the patients within each block. This method of randomization is known as the *permuted-block randomization*, and it is probably the most frequently employed method for the assignment of patients to treatments in clinical trials.

Table 4.3.7 Example of Permutated-Block Randomization with Random Allocation for Four Centers and a Block Size of Four

Random codes for drug XXX, protocol XXX-014
Double-blind, randomized, placebo-control, two parallel groups

A. Hope, MD			J. Smith, MD		
Subject Number	Random Permutation	Treatment Assignment	Subject Number	Random Permutation	Treatment Assignment
1403001	3	Placebo	1401001	1	Active drug
1403002	4	Placebo	1401002	3	Placebo
1403003	1	Active drug	1401003	2	Active drug
1403004	2	Active drug	1401004	4	Placebo
1403005	2	Active drug	1401005	4	Placebo
1403006	4	Placebo	1401006	1	Active drug
1403007	1	Active drug	1401007	2	Active drug
1403008	3	Placebo	1401008	3	Placebo
1403009	2	Active drug	1401009	1	Active drug
1403010	4	Placebo	1401010	4	Placebo
1403011	3	Placebo	1401011	2	Active drug
1403012	1	Active drug	1401012	3	Placebo
1403013	1	Active drug	1401013	4	Placebo
1403014	3	Placebo	1401014	1	Active drug
1403015	2	Active drug	1401015	2	Active drug
1403016	4	Placebo	1401016	3	Placebo
1403017	1	Active drug	1401017	1	Active drug
1403018	3	Placebo	1401018	4	Placebo
1403019	2	Active drug	1401019	2	Active drug
1403020	4	Placebo	1401020	3	Placebo
1403021	3	Placebo	1401021	2	Active drug
1403022	2	Active drug	1401022	4	Placebo
1403023	1	Active drug	1401023	3	Placebo
1403024	4	Placebo	1401024	1	Active drug

To illustrate permuted-block randomization, consider the following example. Suppose that a clinical trial is to be conducted at four centers with 24 patients in each center in order to investigate the effectiveness and safety of a test drug compared to an inert placebo. Also suppose that the permuted-block randomization with a block size of 4 patients is to be employed to prevent treatment and possible covariate imbalances. Two methods can be used to randomly assign patients to treatments. The first method is simply to adopt the method of random allocation within each block by generating a random permutation of numbers 1–4 and assigning the first two to the test drug. This method can be easily implemented using the SAS® procedure PLAN. Table 4.3.7 provides a listing of randomization codes generated by the permuted-block randomization with random allocation within each block.

Since there are two treatments with a block size of 4, we have the following six possible permutations for random assignment of patients to treatments:

1: *TTPP*

2: *PPTT*

Table 4.3.7 Example of Permutated-Block Randomization with Random Allocation for Four Centers and a Block Size of Four (*Continued*)

Random codes for drug XXX, protocol XXX-014
Double-blind, randomized, placebo-control, two parallel groups

M. Dole, MD			C. Price, MD		
Subject Number	Random Permutation	Treatment Assignment	Subject Number	Random Permutation	Treatment Assignment
1402001	1	Active drug	1404001	2	Active drug
1402002	3	Placebo	1404002	3	Placebo
1402003	4	Placebo	1404003	1	Active drug
1402004	2	Active drug	1404004	4	Placebo
1402005	4	Placebo	1404005	2	Active drug
1402006	1	Active drug	1404006	4	Placebo
1402007	2	Active drug	1404007	3	Placebo
1402008	3	Placebo	1404008	1	Active drug
1402009	1	Active drug	1404009	4	Placebo
1402010	3	Placebo	1404010	3	Placebo
1402011	2	Active drug	1404011	1	Active drug
1402012	4	Placebo	1404012	2	Active drug
1402013	1	Active drug	1404013	1	Active drug
1402014	4	Placebo	1404014	4	Placebo
1402015	3	Placebo	1404015	3	Placebo
1402016	2	Active drug	1404016	2	Active drug
1402017	3	Placebo	1404017	1	Active drug
1402018	4	Placebo	1404018	3	Placebo
1402019	1	Active drug	1404019	4	Placebo
1402020	2	Active drug	1404020	2	Active drug
1402021	2	Active drug	1404021	2	Active drug
1402022	1	Active drug	1404022	4	Placebo
1402023	4	Placebo	1404023	1	Active drug
1402024	3	Placebo	1404024	3	Placebo

3: TPTP

4: TPPT

5: PTPT

6: PTTP

where *T* and *P* represent the test drug and the placebo, respectively. Thus there are a total of 6 blocks with 4 patients in each center. The randomization codes can then be generated by producing a random permutation of numbers from 1 to 6, where the numbers correspond to six possible permutations for random assignments of patients as described above. This method is called permuted-block randomization with random selection, and it can also be easily implemented by the SAS® procedure PLAN. Table 4.3.8 provides a listing of randomization codes generated by the permuted-block randomization with random selection. SAS programs for both methods of permuted-block randomization are also provided in Appendix B.3 and Appendix B.4, respectively. It can be verified that the randomization

Table 4.3.8 Example of Permutated-Block Randomization by Random Selection of Blocks for Four Centers and a Block Size of Four

Random codes for drug XXX, protocol XXX-014
Double-blind, randomized, placebo-control, two parallel groups

A. Hope, MD		M. Dole, MD	
Subject Number	Treatment Assignment	Subject Number	Treatment Assignment
1403001	Placebo	1402001	Active drug
1403002	Placebo	1402002	Placebo
1403003	Active drug	1402003	Placebo
1403004	Active drug	1402004	Active drug
1403005	Active drug	1402005	Active drug
1403006	Active drug	1402006	Active drug
1403007	Placebo	1402007	Placebo
1403008	Placebo	1402008	Placebo
1403009	Active drug	1402009	Placebo
1403010	Placebo	1402010	Active drug
1403011	Active drug	1402011	Active drug
1403012	Placebo	1402012	Placebo
1403013	Placebo	1402013	Active drug
1403014	Active drug	1402014	Placebo
1403015	Active drug	1402015	Active drug
1403016	Placebo	1402016	Placebo
1403017	Placebo	1402017	Placebo
1403018	Active drug	1402018	Placebo
1403019	Placebo	1402019	Active drug
1403020	Active drug	1402020	Active drug
1403021	Active drug	1402021	Placebo
1403022	Placebo	1402022	Active drug
1403023	Placebo	1402023	Placebo
1403024	Active drug	1402024	Active drug

codes given in Tables 4.3.7 and 4.3.8 provide treatment balance not only within each study center but also for the entire study.

In addition to the assurance of treatment balance, the permuted-block randomization can account for a possible time-heterogeneous population by forcing a periodic balance. This desirable property of forced periodic balance, however, becomes a disadvantage when the block size is not blinded. In the case where the block size is not blinded, the probability of correctly guessing the treatment increases at the end of each successive block. Matts and Lachin (1988) point out that if the treatment is unblinded, then such forced periodic balance provides investigators an opportunity not only to correctly guess the assignment of patients but also to alter the composition of the treatment groups. As a result, an increasing chance of correctly guessing treatment assignment and alteration of treatment composition can certainly increase the chance of introducing bias to the evaluation of the treatment effect. In addition, as block size increases, the potential selection bias decreases. Although the use of random block size can reduce the selection bias, it cannot completely eliminate the bias. The only way to eliminate the selection bias is to enforce a double-blinded procedure during the entire course of

Table 4.3.8 Example of Permutated-Block Randomization by Random Selection of Blocks for Four Centers and a Block Size of Four (*Continued*)

Random codes for drug XXX, protocol XXX-014
Double-blind, randomized, placebo-control, two parallel groups

J. Smith, MD		C. Price, MD	
Subject Number	Treatment Assignment	Subject Number	Treatment Assignment
1401001	Active drug	1404001	Placebo
1401002	Placebo	1404002	Active drug
1401003	Active drug	1404003	Active drug
1401004	Placebo	1404004	Placebo
1401005	Placebo	1404005	Placebo
1401006	Active drug	1404006	Placebo
1401007	Placebo	1404007	Active drug
1401008	Active drug	1404008	Active drug
1401009	Placebo	1404009	Active drug
1401010	Placebo	1404010	Active drug
1401011	Active drug	1404011	Placebo
1401012	Active drug	1404012	Placebo
1401013	Active drug	1404013	Active drug
1401014	Active drug	1404014	Placebo
1401015	Placebo	1404015	Placebo
1401016	Placebo	1404016	Active drug
1401017	Active drug	1404017	Placebo
1401018	Placebo	1404018	Active drug
1401019	Placebo	1404019	Placebo
1401020	Active drug	1404020	Active drug
1401021	Placebo	1404021	Active drug
1401022	Active drug	1404022	Placebo
1401023	Active drug	1404023	Active drug
1401024	Placebo	1404024	Placebo

study for which both investigators and patients are blinded to block size and treatment assignments. For the method of permuted-block randomization, the variance of the accidental bias, which does not depend on the number of blocks, decreases as block size increases. Although the accidental bias associated with the permuted-block randomization is negligible for large samples, it is more serious as compared to those by simple randomization for small samples.

In summary, ICH E9 guideline suggests that blocking size should be short enough to limit possible imbalance but should be long enough to avoid predictability toward the end of the sequence in a block. Investigator and relevant staff should generally be blinded to the blocking size. In a multicenter trial, it is advisable to have a separate randomization code for each center. In other words, one should apply stratified randomization using center as a stratum or allocate several whole blocks to each center.

It should be noted that permuted-block randomization is also a stratified randomization. Therefore a stratified analysis should be performed with block as a stratum to properly control the overall type I error rate and hence to provide the optimal power for detection of

a possible treatment effect. In practice, the block effect is usually ignored when performing data analysis. Matts and Lachin (1988) show that the test statistic ignoring the block is equal to 1 minus the intrablock correlation coefficient times the test statistic that takes the block into account. Since the patients within the same block are usually more homogeneous than those between blocks, the intrablock correlation coefficient is often positive. As a result the tests that ignore the block will produce more conservative results. However, since most clinical trials are multicenter studies with a moderate block size, it is not clear from their discussion whether a saturated model that factors treatment, center, and block and their corresponding two-factor and three-factor interactions should be included in the analysis of variance. In addition, as given in Tables 4.3.7 and 4.3.8, there are a total of 24 strata (center-by-block combinations) with 4 patients in each stratum. It is not clear whether the Mantel-Haenszel test and linear rank statistics based on the permutation model are adequate for a complete combination of all levels of all strata with a very small number of patients in each stratum. In addition, when the block size is large, there is a high possibility that a moderate number of patients will fall in the last block where the randomization codes are not entirely used up. If the trial is also stratified according to some covariates such as study center, the number of patients in such incomplete block can become sizable. Therefore a stratified analysis can be quite complicated due to these incomplete blocks from all strata and possible presence of an intrablock correlation coefficient.

Adaptive Randomization

As discussed above, the method of complete randomization includes a constant marginal probability for the independent assignment of patients to treatments. However, to some extent it can cause an imbalance in the patient allocation to treatment. On the other hand, the methods of random allocation and permuted-block randomization are useful in forcing a balanced allocation of patients to treatments within either a fixed total sample size or a prespecified block size. These methods of restricted randomization can also maintain a constant marginal probability for the assignment of patients to treatments. In practice, in addition to enforcing a balanced allocation among treatments to some degree, it is also of interest to adjust the probability of assignment of patients to treatments during the study. This type of randomization is called *adaptive randomization* because the probability of the treatment to which a current patient being assigned is adjusted based on the assignment of previous patients. Unlike the other randomization methods described above, the randomization codes based on the method of adaptive randomization cannot be prepared before the study begins. This is because that the randomization process is performed at the time a patient is enrolled in the study, whereas adaptive randomization requires information on previously randomized patients. In clinical trials the method of adaptive randomization is often applied with respect to treatment, covariates, or clinical response. Therefore the adaptive randomization is also known as treatment adaptive randomization, covariate adaptive randomization, or response adaptive randomization. We will now briefly introduce these three applications of adaptive randomization.

Treatment Adaptive Randomization

The treatment adaptive randomization adjusts for the assigning probability of the current patient with respect to the number of patients who have been randomized to each treatment group. Efron (1971) first introduced the idea of biased coin randomization as a method for

adjustment of assigning probability. Consider the same example discussed above. The assigning probability for the first patient is clearly 1/2. After k patients are enrolled, k_T and k_P patients are randomized to the test drug group and the placebo group, respectively. The idea is that if more patients were randomized to the test drug group, then the next patient will be assigned to the placebo group with a probability greater than 1/2. Similarly, if the current number of patients randomized to the test drug group is fewer than that of the placebo group, then the next patient will be assigned to the test drug group with a probability greater than 1/2. If a treatment balance is achieved, then the next patient is assigned to either the test drug group or the placebo group with a probability of 1/2. As an example, Pocock (1984) suggests the use of $p = 3/4, 2/3, 3/5$, and $5/9$, for the chance of less than 5%, for differences in the number of patients between treatment groups being 4, 6, 10, and 16, respectively.

Although the bias coin randomization attempts to achieve a treatment balance by adjusting the assigning probability with respect to the difference in the number of patients who were previously assigned, it may not be satisfactory because a constant assigning probability was used during the entire course of the study. As an alternative, Wei (1977, 1978) consider the so-called *urn randomization*, which is an extension of the biased coin randomization. For the urn randomization, the probability of the assignment of the current patient is a function of the current treatment imbalance (Wei and Lachin, 1988). To illustrate the method of urn randomization, consider a urn that contains exactly A white balls and A black balls. For the assignment of a patient, draw a ball at random from the urn and replace it into the urn. If the drawn ball is a white one, then the patient is assigned to the test drug group. Otherwise, the patient is assigned to the placebo group. Therefore, the assigning probability for the first patient is 1/2. The procedure is to add B white (black) balls to the urn if the drawn ball is a black (white) one. This randomization process is repeated whenever a new patient is enrolled. From the above description, it can be seen that the assigning probability of the urn randomization is determined by A and B . Therefore a urn randomization is usually denoted by $UR(A, B)$. If at each drawing no additional ball is returned to the urn, then the urn randomization is simply a complete randomization and because with replacement, an equal assigning probability of 1/2 is employed for the random selection. If we do not put any ball initially in the urn but use the assigning probability of 1/2 for the first patient, then the subsequent probability of assigning a patient to the test drug after k patients have been enrolled is equal to the proportion of patients who were randomly assigned to the placebo group. This proportion is independent of the number of B balls scheduled to return to the urn.

The urn randomization can achieve a certain degree of a desired balance at the early stage of the study. This is usually accomplished by choosing appropriate numbers of A and B . As pointed out by Lachin, Matts, and Wei (1988), if the ratio of A to B is large, then the urn randomization is very similar to the complete randomization. If the investigator desires to have the treatment balance at the early stage of the study and wishes to maintain a certain degree of balance at the end of the study, then we can choose a large ratio of B to A . This nice property is especially attractive for the post hoc stratified analysis and sequential trial because the size of the post hoc-defined strata and the number of patients at the early termination are usually not known at the planning stage of the trial.

As the sample size increases, the urn randomization approaches complete randomization. As a result, the expected bias factor will be very close to zero. Consequently the selection bias according to the Blackwell-Hodges model will be negligible. For finite samples the selection bias of the urn randomization is smaller than those methods of restricted

randomization, though it can be very close to that of the permuted-block randomization for a sample size fewer than 10. As compared to other methods of randomization, the accidental bias caused by omitting important covariates for estimation of the treatment effect becomes negligible as the sample size increases. However, for small trials an accidental bias may still exist and cannot be ignored. As a result, Lachin, Matts, and Wei (1988) recommend that the urn randomization not be employed for the trial with either the total size or the size of the smallest stratum being fewer than 10. The urn randomization with $A = 0$ and $B = 1$ has nice properties of adequate control for treatment balance. In addition it is less vulnerable to both selection and accidental bias than other methods of restricted randomization. Furthermore, it is easy to implement on a computer because the assigning probability depends only on the current state of treatment allocation.

Although the urn randomization is simple, it requires a much more complicated analysis compared to other methods of randomization. This is because the urn randomization does not have an equal assigning probability for each patient. To conduct an exact permutation test based on the urn randomization, the probability of each assignment is needed in order to compute the p -value. For the large sample size, Wei and Lachin (1988) derive the explicit permutation tests for the logrank and the Peto-Peto-Prentice-Wilcoxon statistics for the censored data. For the urn randomization, statistics of different strata are independent for the prospectively stratified randomization. However, they are correlated for the poststratified subgroup analyses. Wei and Lachin (1988) give an explicit expression for the conduct of a combined test over strata.

Covariate Adaptive Randomization In certain diseases some of the prognostic factors are known to affect clinical outcomes of the treatment. Therefore, it is desired to achieve a covariate balance with respect to these prognostic factors. For this purpose, we may consider to employ the covariate adaptive randomization which is also known as the minimization method (e.g., see Taves, 1974; Pocock, 1984; Spilker, 1991). For an illustration of this method, consider the following hypothetical trial in which a test drug is evaluated with an inert placebo in patients with benign prostatic hyperplasia. Suppose that for patients aged over 64 years old, peak urinary flow rate less than 9 mL/s and an AUA-7 symptom score being at least 20 will have an impact on the clinical evaluation of the test drug. The distribution of these three covariates after 106 patients and 107 patients were enrolled into the placebo and the test drug, respectively, as shown in Table 4.3.9. Suppose that the age of the next patient for randomization is 68 years old with a peak urinary flow rate of 7.4 mL/sec and an AUA-7 symptom score of 21 points. Then one can modify the frequencies of patients with respect to the categories of covariates that this patient falls into. From Table 4.3.9 the numbers of patients who satisfy the criteria (1) age older than 64 years old, (2) peak urinary flow rate less than 9 mL/s, and (3) an AUA-7 symptom score at least 20 for the placebo group are 49, 45, and 29, respectively, while the numbers for the test drug group are 51, 44, and 30, respectively. Therefore, the respective sums for the test drug group and the placebo group are 123 and 125. Since the placebo group has a smaller sum, the procedure is to assign the next patient to the placebo group. Because the minimization method described above is non-random, the covariate adaptive randomization can also use a probability greater than 1/2 to assign the next patient to the treatment group with a smaller sum. Pocock (1984) indicates that assigning a probability of 3/4 or 2/3 may be appropriate. The covariate adaptive randomization requires a constant update of the current status of the covariates. Hence it requires an intensive administrative effort to implement such a procedure even though the computer can alleviate the burden to some extent. As a result, the covariate adaptive randomization

Table 4.3.9 Frequency Distribution of Age, Peak Urinary Flow Rate, and AUA-7 Symptom Score

Covariate	Placebo	Test Drug
<i>N</i>	106	107
Age (years)		
<64	57	56
≥65	49	51
Peak flow rate (mL/s)		
<9	45	44
≥9	61	63
AUA-7 symptom score		
≤7	25	26
8–19	52	51
≥20	29	30

may present a high risk of breaking blindness by either the investigators or the personnel responsible for updating the covariate imbalance status. Note that the covariate imbalance can always be adjusted by a post hoc subgroup analysis when the trial size is moderate. In practice, the covariate adaptive randomization is not recommended for trials with sample sizes greater than 100. More details on the covariate randomization can be found in Pocock and Simon (1975), White and Freedman (1978), and Miller et al. (1980).

Response Adaptive Randomization Another adaptive randomization is to adjust for the assigning probability according to the success or failure of the treatments to which previous patients were assigned. This idea was first proposed by Zelen (1969) and subsequently known as the play-the-winner (PW) rule. For the first patient enrolled in the study, an assigning probability of 1/2 is employed to either treatment. Suppose that the white ball represents the test drug (*T*) and the black ball represents the placebo (*P*). If the current patient receives treatment *T* and the response is a success or if the current patient receives *P* and the response is a failure, then put a white ball in the urn. If the current patient receives *T* and the response is a failure or if the current patient receives *P* and the response is a success, then put a black ball in the urn. When the next patient is enrolled into the trial, we randomly draw a ball without replacement from the urn. If there is no ball in the urn, then an assigning probability of 1/2 is employed to either treatment. Wei and Durham (1978) indicate that the responses of patients might not be observed before the arrival of the next patient. The urn therefore might have a very high possibility of being empty during the entire course of the trial. It turns out that the assigning probability of patients is approximately 1/2 under the play-the-winner rule which is quite similar to the method of random allocation. If the response is unavailable in a short period of time before the next patient is enrolled, Zelen (1969) suggests that one can continue to assign the same treatment if the response of the current patient is a success but switch to the other treatment if a failure is observed. This rule is called the *modified play-the-winner* (MPW) rule. Note that the MPW rule is a deterministic rather than stochastic process.

To overcome the drawback of a deterministic process, Wei and Durham (1978) suggest an alternative method known as the *randomized play-the-winner* (RPW) rule. At the beginning of the trial, an equal number (*m*) of white and black balls are placed in an urn, where the white balls represent the test drug and black balls represent the placebo. When a patient

is enrolled into the trial, a ball is drawn at random from the urn with replacement. If the randomly selected ball is a white one, the patient is assigned to the test drug group, and otherwise, to the placebo group. If the previous patient was assigned to the test drug group and the response is a success, then additional B white balls and A black balls are put into the urn, where $B \geq A \geq 0$. If the response of the previous patient receiving the test drug is a failure, then additional A white balls and B black balls are put into the urn. Similarly, if the previous patient was assigned to the placebo group and the response is a success, then additional A white balls and B black balls are put into the urn. If the response of the previous patient receiving the placebo is a failure, then additional B white balls and A black balls are put into the urn. If the urn is empty, then a probability of $1/2$ is used. It should be noted that exactly additional A plus B balls are put into the urn whenever a response is available for the assignment of the next patient. The RPW does not require the availability of the response of the previous patients. The RPW, which is random, provides a higher probability to treat the next patient, with the better treatment based on the current result of the trial. When $A = B$, RPW is the same as the method of complete randomization. Wei and Durham (1978) show that when the ratio of B to A becomes large, the probability of assigning patients to the test drug is approximately equal to the ratio of the failure rate of the placebo to the sum of failure rates for both treatments. Therefore, if the ratio of B to A is large, then the RPW tends to assign more patients to the better treatment. In addition, Wei et al. (1990) indicate that RPW is less vulnerable to the experimental bias than other adaptive randomizations.

Since the assigning probability of the current patient to treatments adjusts for the past history of the outcomes of the previously randomized patients, the statistical analysis based on RPW is much more complicated than those based on other methods of randomization. Wei (1978) describes the permutation distribution of a test for the binary response under RPW. In addition, Wei et al. (1990) study the exact conditional, exact unconditional, and approximate confidence intervals for the treatment difference in binary responses. The results indicate that the exact unconditional procedure performs much better than the conditional procedure. In addition the large sample unconditional confidence intervals derived from the likelihood statistic are not very sensitive to the adaptive randomization and perform quite satisfactory for trials with moderate sample size. The confidence intervals based on the maximum likelihood estimates behave very poorly under RPW. Therefore, Wei et al. (1990) suggest that the features of response adaptive randomization be taken into account in the analysis. On the other hand, Tamura et al. (1994) perform a Bayesian analysis for a trial concerning patients with depressive disorder using RPW.

Recently several clinical trials were conducted using the play-the-winner rule. For example, a clinical trial was conducted at the University of Michigan to investigate the effectiveness and safety of extracorporeal membrane oxygenation (ECMO) in treating newborn babies with persistent pulmonary hypertension (PPH) with the conventional mechanical ventilation (CMV) as the concurrent control (Cornell et al., 1986). Past experience has shown that infants with PPH has a 80% death rate in the absence of ECMO which is an artificial heart-lung machine recycling the blood through a membrane exposed to the oxygen with a high concentration. On the other hand, ECMO is a surgical procedure with potential life-threatening complications to the infants. Since the response (either death or recovery) can be observed within a few days, the response of the previously treated infants is available before the entry of the next newborn. Consequently, a RPW with $m = 1$, $A = 0$, and $B = 1$ was employed for the first infant. The result turned out to be the ECMO treatment, and the baby recovered. Then, a white ball representing the ECMO treatment was put in the urn. For

the next infant, although the probability of assigning infants to the ECMO treatment is 2/3 and to the CMV treatment is 1/3, the resulting treatment for the second infant is the CMV treatment with an unfortunate outcome of death. Therefore a black ball representing the CMV treatment was put back into the urn. This procedure was employed. The result turned out that the ECMO treatment was randomly assigned to the next eight consecutive infants all of whom survived. The trial was terminated at this point. Note that two more infants were also assigned to the ECMO treatment without invoking the ECMO treatment, but both of them survived too. Boston's Children Hospital Medical Center and Brigham and Women's Hospital conducted a similar adaptive trial to compare the ECMO treatment with the CMV treatment in infants with PPH (Ware, 1989). The study involved two phases. For the first phase, a permuted-block randomization with a block size of 4 was used to generate randomization codes for the trial. It was calculated at the planning stage that if 4 deaths were observed in one of the two groups, this phase would be terminated, and the study would proceed to the second phase in which all subsequent infants would be assigned to receive the other treatment. It was also predetermined that the second phase would be terminated if 4 deaths or 28 survivors were observed from the infants of both phases who were enrolled into the other treatment. For the first phases of the trial with a permuted-block randomization, 10 infants were assigned to the CMV treatment and 9 to the ECMO treatment. Four infants who were randomly assigned to the CMV treatment died, and all 9 infants receiving the ECMO treatment during the first phase survived. As a result all subsequent infants were assigned without randomization to receive the ECMO treatment. The trial was terminated at the 20th infant who enrolled into the second phase and did not survive.

For another example, Tamura et al. (1994) report that a clinical trial utilizing the RPW rule was conducted to assess the efficacy and safety of fluoxetine as compared with a placebo in patients with depressive disorder. The study was stratified according to rapid eye movement latency (REML) which is defined as the time between sleep onset and the first rapid eye movement. If REML of patients is shorter than or equal to 65 minutes, then he or she is stratified into the shortened REML; otherwise, he or she is stratified to the normal REML group. A patient is classified as a responder if the percent reduction at the final eight-week visit from baseline on the first 17 items of the Hamilton Depression scale (HAM-D-17) is at least 50%. Since a period of eight weeks is required to observe the primary endpoint and the patient accrual was rather rapid, this time delay could not allow the investigator to employ the RPW with the primary endpoint. As a result a surrogate endpoint of the percent reduction of at least 50% in HAM-D-17 in two consecutive visits after at least three weeks of therapy was used for the response adaptive randomization. Within each stratum the first six patients were assigned using the method of permuted-block randomization. Starting with the seventh patient, the RPW was initiated within each stratum with $m = 1$, $A = 0$, $B = 1$. A total of 89 patients were randomized, and yet the surrogate endpoint was only observed in 61 of the 89 randomized patients and 83 patients were included in the analysis based on the primary endpoint. Tamura et al. (1994) indicate that their experience with RPW for this trial has been generally positive despite increasing communication between the sponsor and investigators.

As indicated earlier, statistical inference depends on the statistical test used, which in turn depends on the randomization employed. It is therefore important to derive an appropriate statistical test according to the randomization employed. For example, the ECMO study conducted at the University of Michigan created a controversy over the statistical analysis used for comparing the two treatments. Recall that the 11 infants assigned to the ECMO treatment survived, and only one baby, the one assigned to the CMV treatment,

died. Although the ECMO trial at the University of Michigan only involved 12 infants, the results of this study have raised many serious questions regarding complicated statistical and ethical issues. First, how does one compare two treatments with such a severe treatment imbalance (i.e., 11-ECMO versus 1-CMV)? A sample of one patient contributes very little information toward the comparison between treatments. Second, there exists no appropriate statistical test under the RPW model. Alternatively, Cornell, et al. (1986) considered the method of ranking and selection to demonstrate that the ECMO treatment is superior to the CMV treatment. However, his method does not provide *p*-values and confidence intervals. Wei (1988) developed a permutation test under the RPW rule. However, Begg (1990) pointed out that Wei's permutation test is inappropriate and obtained some *p*-values (ranging from 0.038 to 0.62) based on different analyses (also see the discussion by Wei, 1990; Pocock, 1990; Cox, 1990). For the ECMO study conducted by Harvard Medical School, there was employed a two-stage design. At the first stage, a probability of 1/2 was used to assign infants until there was statistically significant evidence that one treatment showed a superior efficacy; then all remaining infants were assigned to the superior treatment (Ware, 1989). This example demonstrates that appropriate statistical procedure must be derived for the method of randomization to be employed in clinical trials.

For the response adaptive randomization, despite its advantage in ethical terms, it is not widely accepted in clinical trials (Simon, 1991; Rosenberger and Lachin, 1993; Rosenberger, 1999). This is probably due to the availability of appropriate statistical tests under various methods of response adaptive randomization. Rosenburger and Lachin (1993), however, provide a list of general conditions under which a response adaptive randomization can be implemented successfully given the existing methodology. These conditions are summarized below:

1. There is a single outcome or hypothesis of interest.
2. Outcomes are ascertainable in a short period of time.
3. The study has important public health consequences, but the diseases are not life-threatening.
4. The study has an adequate sample size and the composition of the sample is not likely to change over time.
5. The participants in the study have the resources to logically implement the randomization procedure.

These general conditions may limit the application of the response adaptive randomization to clinical trials. First, a disease is a medical condition that is very complicated and usually cannot be adequately described by a single clinical outcome. Hence a clinical trial may have more than one objective based on more than one clinical outcome. Consequently the first condition seems to be very difficult to be satisfied by most clinical trials. Second, under the RPW model, despite recent developed analysis procedures for binary data (Wei, 1988), multinomial and continuous data for large samples (Rosenberger, 1993, 1999), analyses for the secondary clinical endpoints, censored data and subgroup analyses for the adjustment of covariates have not yet been fully developed. Finally clinical trials usually require multiple visits for an evaluation of the treatment's progress. However, statistical method for the analysis of repeated measurements has not been proposed for the RPW rule. As a result Rosenberger and Lachin (1993) and Rosenberger (1999) conclude that the future use of the response adaptive randomization is uncertain.

4.4 IMPLEMENTATION OF RANDOMIZATION

In the pharmaceutical industry, for good clinical practice a set of standard operating procedures (SOP) for generation, implementation, and administration of randomization is usually established to ensure the integrity of clinical trials. In this section we will introduce an implementation procedure for the method of nonadaptive randomizations which is adopted by most of pharmaceutical companies for clinical research and development. In the pharmaceutical industry the department of Biostatistics and Data Management (or Biometrics) is usually responsible for the activities of statistics, programming, and clinical data management. Within the department, a drug-specific (or project-specific) team (or unit) is usually formed to oversee the development of statistics, programming, and data management during the process. This team usually consists of biostatisticians, programmers, and data coordinators. Note that unlike a clinical research associate or monitor, this team does not involve itself with the day-to-day activities of clinical projects. However, this team is responsible for the selection of randomization methods, case report forms design and review, clinical data management, statistical analysis, and report writing. Because generation of randomization codes is the key to the success of the intended trials, a group within the department of Biostatistics and Data Management, not involved with clinical trials, is designated to be responsible for generation and management of randomization codes. Since there are many different methods for generating randomization codes as discussed above, the group is responsible for the implementation of a system that incorporates the various methods of randomization by developing computer programs. Since the system, which may contain a number of computer programs, is not designed for commercial but for internal use, it is recommended that the methods of randomization employed and the corresponding computer programs be adequately documented. Also it is desirable for a user-friendly User's Reference Manual to be developed. The User's Reference Manual should contain detailed instructions for the use of the system, references to the pseudonumber generator, methods of randomization, programs for listings of the pseudonumber generator and for the production of a listing of the randomization codes. In addition a prospective validation of the design programs should be performed according to a validation protocol before the implementation of the system. Note that the FDA requires that the results of the validation test be documented and that the system be validated periodically.

Generation, Labeling, and Packaging

During the development of the clinical protocol, the project clinician and biostatistician usually discuss the selection of an appropriate method of randomization and some related logistic issues for the implementation of randomization according to study objectives, primary endpoints, stratified covariates (if any), and sample size of the trial. The randomization method employed for the study should be described in detail in the study protocol without disclosure of the block size, if the permuted-block randomization is used. The study protocol and the investigators's brochure should also describe in detail a standard procedure for treatment assignment and drug dispensing. In general, patients should not receive any medication unless they have met all eligibility criteria and have signed the informed consent forms as defined in the study protocol. A formal request for randomization codes cannot be sent to the project statistician unless the study protocol has obtained an approval from an internal protocol review committee. The project statistician can then check whether the request is adequate with respect to the study protocol and design. The

randomization codes will be generated according to the selected method of randomization by the randomization group if no concerns are raised by the project statistician. The group is not only responsible for the generation of randomization codes but also for performing quality assurance (QA) procedures of the generated randomization codes. The QA procedures are to (1) check every generated randomization codes, (2) document the program logs for generation of the randomization codes, and (3) maintain information for the generation of the randomization codes including the seed and the first and last random numbers generated from the seed. If the randomization codes meet the requirements of the QA procedures, then a list of randomization codes is sent to the drug packaging department or some contracted laboratory for packaging the study drugs. If the trial is a triple-blind study, the project statistician and clinician or other project team members should be informed only of the generation of the randomization codes by a cover memo. If, however, the trial is a double-blind study, then the project statistician and clinician might get a copy of randomization codes upon request. The information of the randomization codes will then be locked in the database until the time at which an interim analysis or final analysis is performed. For a triple-blind study the clinical data coordinator and clinical research associate identify the patients through a sequentially assigned patient (subject) number to maintain the blindness. A patient or subject number usually contains three parts, which include the project number, the study center number, and a sequentially assigned patient number within the individual study center. Let us take, for example, a number 01401015 (i.e., 014-01-015) used to identify a patient in a clinical trial. The first three digits 014 are an identifier of the study drug XXX, the next two digits 01 represent the first study site, and the last three digits 015 indicate that the patient is the 15th patient to enroll in the study. The project team will also generate a set of dummy randomization codes for the project statistician to perform necessary programming for patient listings or case report tabulations as required by the FDA to shorten the statistical analysis after the study is completed and the database is locked.

When the drug packaging department receives the randomization codes, the study drugs are packed according to the method and instruction as stated in the protocol. The most secure method for maintaining blindness is to use identical blister packs or drug kits with identically appearing contents. Usually the drug kits have a three-part double-blind tear-off label affixed to the cover of the kit. This label has the protocol number and the preprinted patient number. Patients' initials and the time and date the drug dispensed will be recorded on each label. The time and date are important for establishing an audit trail of treatment assignment. The double-blinded tear-off portion, which will be attached to the appropriate page in the case report form, contains the actual treatment group information to which the patient is assigned. These sealed labels will not be opened unless it is required in a medical emergency when knowledge of the respective treatment may influence medical care. At the conclusion of the study, the investigators should return all used and unused study drugs to the sponsor. Usually there is a boilerplate paragraph included in the study protocol for drug accountability. Figures 4.4.1 and 4.4.2 provide flow charts of the randomization procedure discussed above.

Random Assignment

In the pharmaceutical industry, the randomization procedure is not limited to the generation of randomization codes for treatment assignments. It can also be applied to laboratory evaluations. For example, routine hematology, blood chemistry, urinalysis, or some other

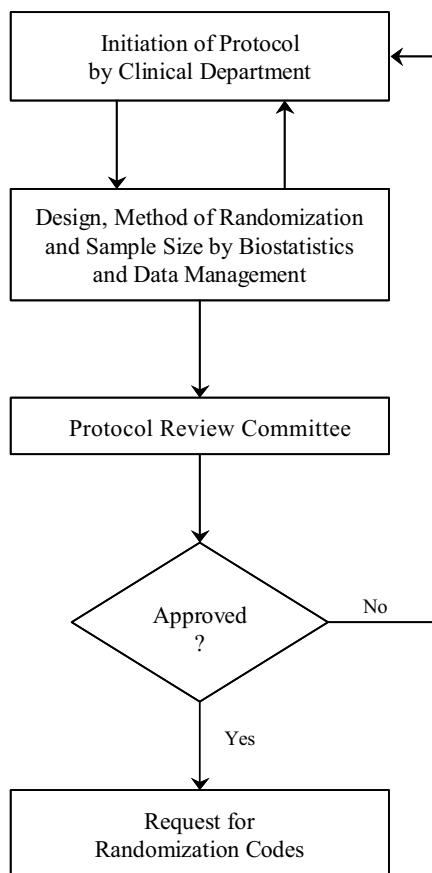
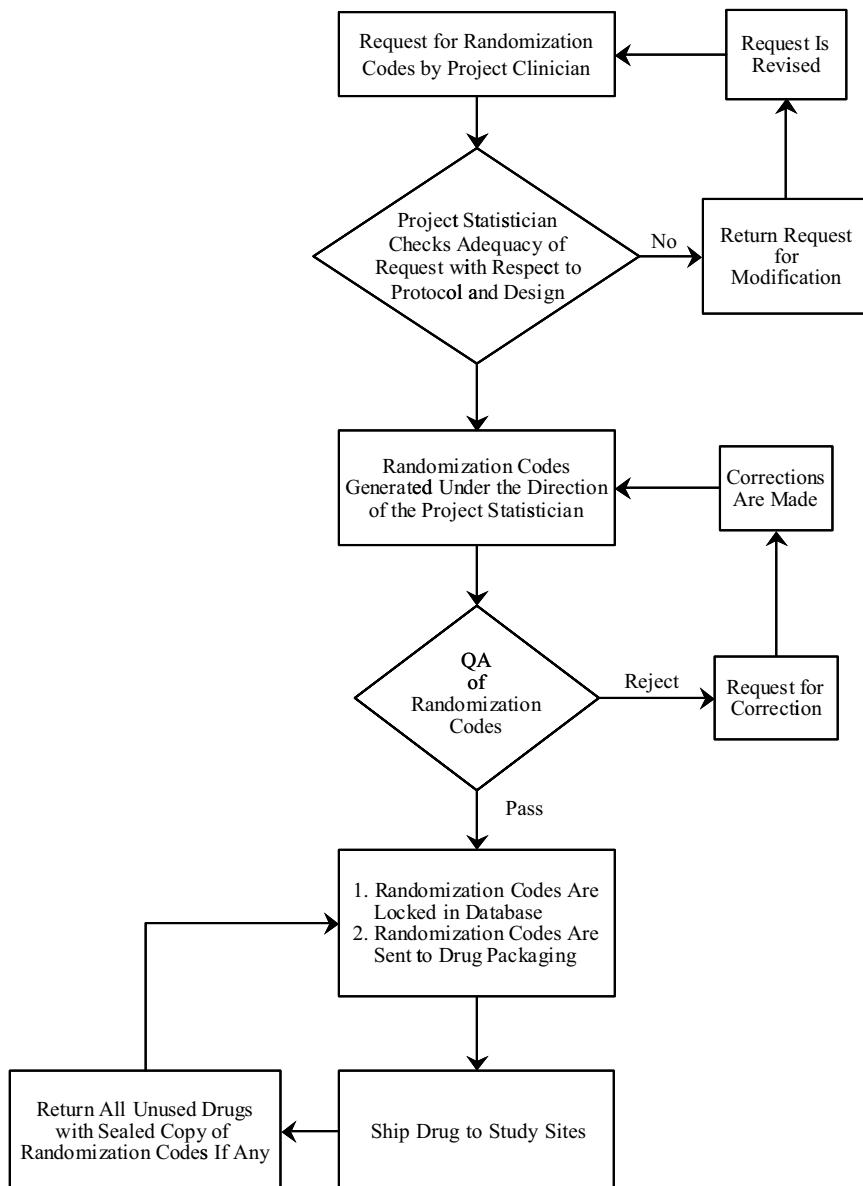


Figure 4.4.1 Randomization codes at the planning stage.

special compounds such as serum hormone levels of the patients from a clinical trial are usually assayed at a centralized contracted laboratory. However, the assay of active ingredients of the study drugs are often performed by the method developed by the laboratory. Since samples from a clinical trial may be enormous due to multiple visits, assays may be required to perform at different times of a day over a period of several days due to the capacity of the laboratory. As a result, a proper design for the drug assay is necessary to eliminate variability due to analyst, time, and day. In addition the assay should be performed in a blinded fashion to avoid possible bias caused by the knowledge of the study drugs. Therefore randomization codes may also be generated by the same randomization group according to the design and method of randomization as deemed appropriate by the project statistician. The randomization codes for treatment assignments and drug assays are to be stored in the central file and cannot be released until the database is locked. The generation and implementation of randomization codes for drug assays can be logically complicated when assays for different active ingredients or their metabolites are required to perform individually with subdivisions of the blood samples.

In some cases, randomization is also used for poststudy evaluation. For example, for the evaluation of contrast agents in the enhancement of images obtained with magnetic

**Figure 4.4.2** Generation of randomization codes.

resonance imaging (MRI), two sets of films with and without contrast agent are usually obtained. The set of films without the contrast agent is obtained before that with the contrast agent. In general, this type of trial is conducted in an open-label fashion without a concurrent control because it is almost impossible to maintain the blindness during the trial. As a result the FDA requests that the sponsors perform a blinded reader study after the trial is completed. A separate protocol for the blinded reader study is also prepared after all films are obtained from the clinical trial. The blinded reader studies are considered

as adequate well-controlled studies for approval. The package insert is usually derived from the blinded reader studies rather than the actual clinical trials. For blinded reader studies, several qualified readers, who are not engaged the clinical part of the trial and are not associated with any investigators of the trial, are asked in a random but blinded fashion to evaluate the films. For this purpose, randomization codes for a blinded reader study can also be generated according to the design and randomization method specified in the protocol under the supervision of the project team.

Note that the randomization procedure described above is probably the most frequently employed procedure for conducting clinical trials in the pharmaceutical industry with sample sizes smaller than a few thousands. For most clinical trials sponsored by the NIH or other cooperative groups, the sample size can be quite large. For example, the ISIS-2 (1988) study randomized 17,187 patients with suspected acute myocardial infarction to four treatment groups and 22,071 male physicians were enrolled to receive one of the four treatments in the U.S. Physician's Health Study (1989). An even larger study is the GUSTO (1993) study which enrolled a total of 41,021 patients with evolving myocardial infarction. As a result, it may not be feasible to adopt the randomization procedure described above. As an alternative, a centralized randomization center may be established for random assignment of treatments either by mail or by telephone. If the time between the screening and request for random assignment is long, say a month, then the mailing system may be possible. For example, see the study conducted by the Coronary Drug Project Research Group (1973) which is described in detail in Meinert (1986). It should be noted that it is not an easy task to handle treatment assignments of more than tens of thousands of patients, especially when the time from the onset of symptom to the treatment is also considered as a crucial factor such as rt-PA for acute ischemic stroke (National Institute of Neurological Disorder and Stroke rt-PA Stroke Study Group, 1995). Hence a central administrated telephone-based assignment system such as Interactive Voice Response System (IVRS, Chen, 2003), should be employed. For example, ISIS-2 (1988) used a 24-hour telephone service, based in Gent and Brussels for Belgium, Berlin for Germany, Valencia for Spain, Bellinzona for Austria and Switzerland, Lyon for France, and Oxford for England and all other countries. The information of patient identifiers such as age, systolic blood pressure, hours from onset of the episode of pain that led to admission, aspirin use during the week before, and the planned treatment in hospital must be completed before a patient is randomized to receive treatments. In addition the method of minimization randomization was also adopted at Oxford for balancing the prognostic factors recorded at entry. However, on January 24, 1986, a programming error was discovered that led more patients randomized at Oxford to being allocated to the placebo infusion and placebo tablets over a period of two months (see Chapter 2 for more information on the treatments of ISIS-2). This programming error was corrected and the exact balance restored in August 1986. Similarly the GUSTO trial used a 24-hour a day, seven-day-per-week randomization center to verify patient eligibility, informed consent, and to assign treatments to more than forty thousand patients. Note that randomization can be performed through a computer networking system such as internet, or web-based networking system. A computerized standard form of eligibility information and informed consent must be sent with the request to the randomization center. Then a validated computer program at the randomization center can immediately enter the data of eligibility for a patient interactively on line through web-interface and verify the patient's eligibility. If inclusion criteria are met and none of the exclusion criteria are observed, then a random assignment of the patient to a particular treatment can be issued in a blinded fashion and sent to the study

center. Otherwise, a message of reasons for refusal to issue randomization codes should be sent to the study center. This randomization process is not only accomplished in seconds but also eliminates the human errors that often occur during the randomization process. It should be noted that a computer system should be validated if it is to be employed for the generation of randomization codes. In addition all personnel should have appropriate training. It is suggested that several dry runs with simulated cases be done before the actual implementation of the system takes place.

4.5 GENERALIZATION OF CONTROLLED RANDOMIZED TRIALS

In most clinical trials the group of patients (or sample) who participate is just a small portion of a heterogeneous patient population with the intended disease. As indicated earlier, a well-controlled randomized clinical trial is necessary to provide an unbiased and valid assessment of the study medicine. A well-controlled randomized trial is conducted under well-controlled experimental conditions, which are usually very different from a physician's best clinical practice. Therefore it is a concern whether the clinical results observed from the well-controlled randomized clinical trial can be applied on the patient population with the disease. As a result the feasibility and generalization of well-controlled randomized trials have become an important issue in public health (Rubins, 1994). For illustration purposes, consider the following two examples.

In early 1970s, a high cholesterol level was known to be a risk factor for developing coronary heart disease. To confirm this, a trial known as the *Lipids Research Clinics Coronary Primary Prevention Trials* (CPPT) was initiated by the National Heart, Lung, and Blood Institute to test the hypothesis whether lowering cholesterol can prevent the development of coronary heart disease. In the CPPT trial, a total of 4000 healthy, middle-age males were randomized to receive either the cholesterol-lowering agent cholestyramine or its matching placebo (Lipids Research Clinics Program, 1984). The primary endpoint was the incidence of coronary heart disease after a seven-year follow-up. A statistically significant reduction of 1.7% in 7-year incidence of coronary heart diseases was observed for the cholestyramine group as compared to the placebo (8.1% versus 9.8%). An expert panel recommended to extrapolate the results for the treatment of high cholesterol in populations that had never been studied and whose benefit has not yet been demonstrated (The Expert Panel, 1989; Recommendations for the Treatment of Hypercholesterolemia, 1984). Moore (1989), however, raises a serious doubt regarding the expert panel's recommendation for the treatment of patients with high cholesterol levels. Moore points out that the CPPT trial was conducted on middle-age males which cannot be applied to a general patient population with hypercholesterolemia. Another example concerning the generation of controlled randomized trials is the U.S. Physician's Health Study described earlier. The question is whether the benefit regarding fatal and nonfatal coronary heart disease, which was observed using 22,000 highly educated males aged over 40 years old, can also be observed in an average individual regardless of gender, race education, and socioeconomic background. This question is indeed a tough one to answer. We can address the question in part by performing a subgroup analysis with respect to the composition of the patients in the trial. This study led to the United States Congress passing legislation (National Institute of Health Reauthorization Bill, 1993) which requires the specification of the composition of any human studies sponsored by the NIH. More detail can be found in Wittes (1994).

One way to ensure the generalization of controlled randomized trials is to understand the process for drawing statistical and clinical inference. Basically statistical and clinical inference for the generalization of results obtained from clinical trials to other patients is a two-step process. The first step is to *internally* apply the statistical and clinical inference on the targeted population to other patients within the population. The second step is to *externally* generalize the statistical/clinical inference made on the targeted population to another patient population with different characteristics. These steps involve the concept of *population efficacy* (or safety), *individual efficacy* (or safety), reproducibility and generalizability which will be illustrated below.

Note that the current conduct of clinical trials is to compare the difference in distributions of the clinical responses observed from patients under a test therapy and a standard (or reference) therapy or a placebo. This concept is referred to as population efficacy (or safety). Suppose that the distribution of a clinical response can be adequately described by a normal probability distribution. Then the population efficacy can be assessed through the comparison of the first two moments of the distributions between the test and the reference therapies. This is because a normal distribution is uniquely determined by its first two moments. The comparison of the first moment of the efficacy endpoints for the two therapies is usually referred to as *average efficacy*, while the comparison of the second moments is called the *variability of efficacy*. To provide a better understanding of average efficacy and variability of efficacy, the comparison in averages and variabilities are illustrated in Figures 4.5.1 through 4.5.3. For example, to compare the reduction in diastolic blood pressure for evaluation of a new antihypertensive agent against a placebo, Figure 4.5.1 shows that the two distributions are very close in both average and variability, which indicates that there is no difference in average and variability of the reduction of diastolic blood pressure. Therefore the new agent may not be efficacious. On the other hand, Figure 4.5.2 demonstrates that the new agent is more effective in reducing blood pressure. Note that in most clinical trials with continuous primary endpoints, the objectives are often formulated as hypotheses for testing the average efficacy. As a result, the population efficacy of the new therapy is often

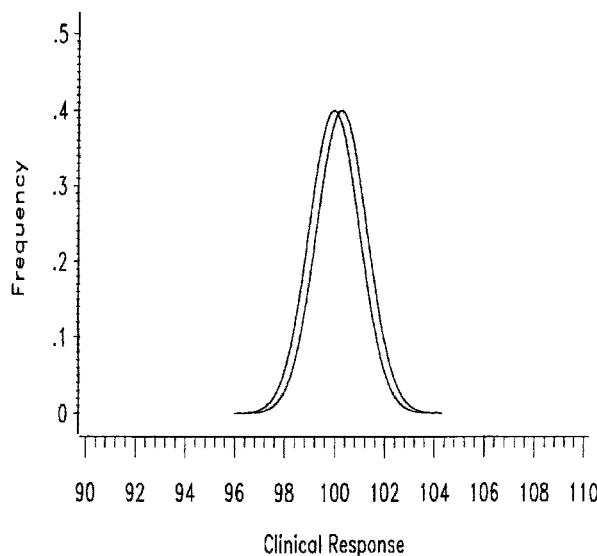


Figure 4.5.1 Population efficacy in averages and variabilities.

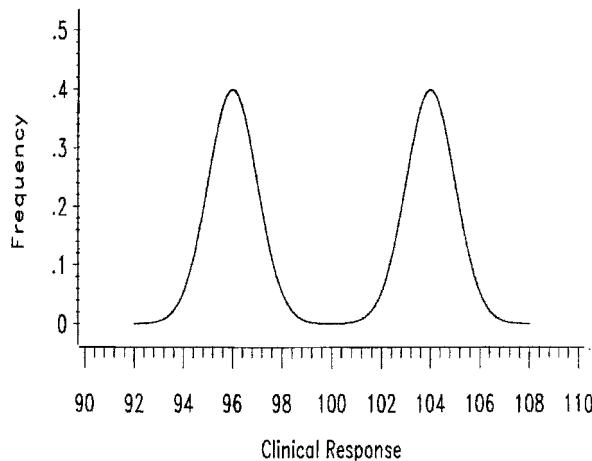


Figure 4.5.2 Population efficacy in averages and variabilities. Unequal averages and equal variabilities.

assessed through the average efficacy under the assumption of equal variability of efficacy. This assumption, which should be verified, is often ignored by both clinicians and biostatisticians. As illustrated in Figure 4.5.3, it is not uncommon that the new agent shows a better efficacy than the placebo and yet exhibits a much larger variability. Since the large variability of the new agent may cause a safety concern, it is recommended that the possible causes of the large variability be carefully examined. A large variability may be due to differences in the composition of patients such as biological variation between two populations. This will certainly have an impact on the generalization of the results to other populations. For population efficacy (or safety), we might first generalize the results to similar but slightly different populations and then, in stages, to much different populations. This concept of generalization is illustrated in Figure 4.5.4 as similarity circles. The strength of the

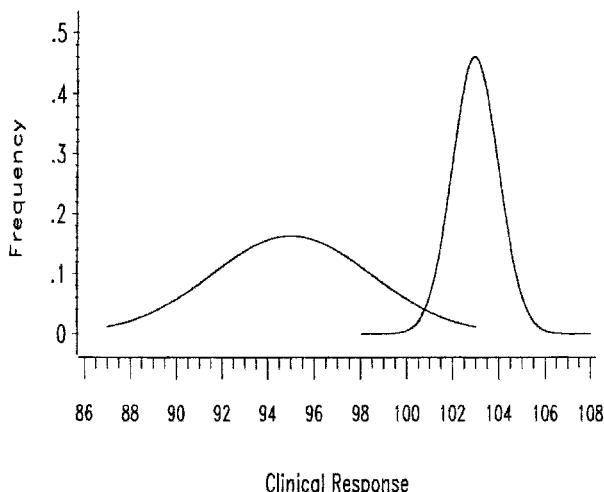


Figure 4.5.3 Population efficacy in averages and variabilities. Superior average efficacy and unequal variabilities.

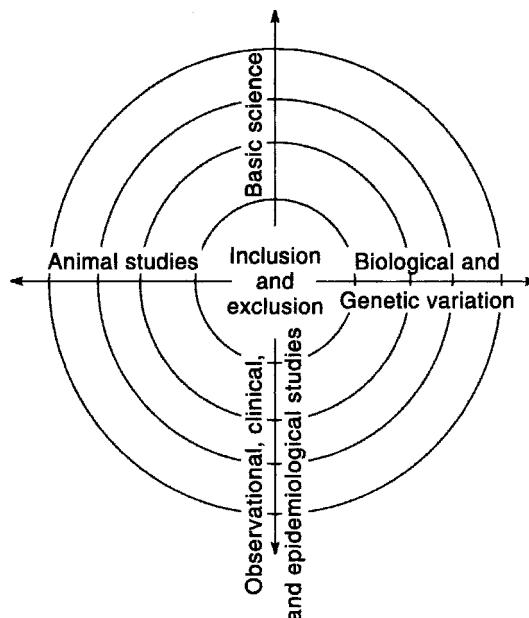


Figure 4.5.4 Generalization of clinical results as similarity circle.

generalization is assessed by the distance between any two points within the circle. Note that the distance is a measure of similarity between populations, which is a function of factors such as basic science, animal models, biological variation, and results from other types of studies.

Note that the establishment of population efficacy does not guarantee that the results can be generalized to a patient with his or her own biological and genetic makeup, educational status, and socioeconomic status who is cared by a particular physician at a different geographical location. The reason is that the efficacy is not established within the patient. The concept for the comparison between the two distributions of the primary efficacy (or safety) endpoints obtained from the same patient under repeated administrations of the new agent and the reference is called *individual efficacy* (or *safety*). The concept of individual efficacy is not new and has been advocated by many clinical researchers. See, for example, Guyatt et al. (1986) and Sackett (1989). Guyatt et al. (1986) attempt to evaluate individual efficacy of theophylline through a *N-of-1* randomized trial concerning a patient with asthma. The *N-of-1* randomized trial was conducted based on the following assumptions and procedures: First, the patient and his or her attending physician determined symptoms such as shortness of breath on ordinarily daily activities, nocturnal spasms of dyspnea, and coughing as primary clinical responses for the treatment of theophylline. The patient agreed to record standardized measures of severity of these symptoms. It was also decided that a 10-day treatment would be long enough to evaluate the effectiveness of the treatments. The *N-of-1* randomized trial was performed in a double-blind fashion with randomization of the order of treatments. At the end of each pair of treatment periods, the patient and physician met to examine the results (also in a blinded fashion) and decided whether to stop or to continue another pair of treatments. After administration of two pairs of treatments in a blinded and random fashion, the analysis detected a statistically significant difference between the

treatments. When the randomization codes were unblinded, it was found that the patient was better on placebo than theophylline.

Repeated administrations of the test and reference therapies within the same patients made the comparison between distributions of the primary clinical end-points within the same individual possible. If we perform this type of trial over N patients in a similar manner, then a total of N pairs of distributions of the test and reference therapies can be generated. Consequently both population and individual efficacy can be made based on these N pairs of distributions. First, within each individual, the individual average efficacy of the test therapy is assessed as the difference between averages of two distributions. In addition the individual variability of efficacy can be evaluated as the ratio of individual intrapatient variabilities between the distributions obtained under the two treatments from the same patient. As a result individual efficacy can be evaluated by comparing averages and variability of the two distributions obtained from the same patient. Since the individual average efficacy and variability of efficacy are obtained from all N patients, we can perform a statistical test to see whether the individual average efficacy and individual intrapatient variability are homogeneous across these N patients. The concept of homogeneity of individual average efficacy and individual intrapatient variability is referred to as patient-by-treatment interaction for average and variability, respectively. A patient-by-treatment interaction implies that the relative efficacy of the test therapy varies from patient to patient. Therefore, if a patient-by-treatment interaction is found, then the relative efficacy of the test agent must be assessed individually for each patient, that is, the individual efficacy. On the other hand, if the relative efficacy is not heterogeneous, then the information of individual average efficacy and individual intrasubject variability can be combined over N patients to provide a basis for population efficacy. The concepts of population and individual efficacy are motivated from population and individual bioequivalence (e.g., see Chow and Liu, 1995b, 2000); they are important concepts for evaluation of bioequivalence between a brand-name drug product and its generic copies. However, the concept of individual efficacy (safety) has not been accepted by nor has convinced the clinical/medical community. Guyatt et al. (1986) point out that the limitations of individual efficacy include (1) it cannot be applied to a disease that can be cured in a short period of time, and (2) it cannot be assessed with the hard clinical endpoints such as death or other irreversible condition indicators.

4.6 BLINDING

Although the concept of randomization is to prevent bias from a statistically sound assessment of the study drug, it does not guarantee that there will be no bias caused by subjective judgment in reporting, evaluation, data processing, and statistical analysis due to the knowledge of the identity of the treatments. Since this subjective and judgmental bias is directly or indirectly related to treatment, it can seriously distort statistical inference on the treatment effect. In practice, it is extremely difficult to quantitatively assess such bias and its impact on the assessment of the treatment effect. In clinical trials it is therefore imperative to eliminate such bias by blocking the identity of treatments. Such an approach is referred to as *blinding*. Blinding is defined as an experimental condition in which various groups of the individuals involved with the trial are withheld from the knowledge of the treatments assigned to patients and corresponding relevant information. The blinding is also known as *masking* by some research organizations such as NIH.

For a clinical trial, if the sponsor is to monitor the study and to perform in-house data management and statistical analysis, then the clinical trial typically involves three parties: the patient, the study center or investigator, and the sponsor. The patient is the most important participant in the clinical trial. No clinical trial is possible without the patient's dedicated participation, endurance, corporation, and sacrifice. The study center, in a broad sense, is referred to as those individuals who are either directly in contact with the patient or perform various evaluations for the patient. Among these individuals is the investigator, who usually is the patient's primary care physician and members of the patient's care team such as the pathologist for histopathological evaluation, the radiologist for imaging assessment, a staff nurse who may also serve as the coordinator for the study center, the pharmacist who dispenses the study medicines, and other health care personnel at the study center including the laboratory staff and the contracted houses that perform the various laboratory evaluations for the blood or urine samples collected from the patient. Note that in clinical trials sometimes the term *investigator* may be used exchangeably with the term *study center* or *study site* in a broader sense. For a clinical trial two functional teams are usually formed by the sponsor. The first team is the clinical/medical team which consists of the project clinician (e.g., physician monitor) and clinical monitor such as the CRA. The project clinician has an overall responsibility for the success of the trial, while the responsibility of the clinical monitor is not only to monitor the conduct of the trial but also to ensure that the investigator adheres to the study protocol. The second team is the biostatistics and data management team which includes the project statistician, programmer, and data coordinator. The project statistician oversees the activities of data management, programming, and statistical aspects of the trial, while the programmer is responsible for programming support for data management, analysis, and report. The data coordinator will coordinate the activity of database setup, data entry, data verification, data query generation/resolution, database cleanup and finalization to ensure the quality of the final database.

Basically blinding in clinical trials can be classified into four types: open label, single blind, double blind, and triple blind. An open-label study is a clinical trial in which no blinding is employed. That is, both the investigator and the patient have an idea about which treatment the patient receives. Since patients may psychologically react in favor of the treatments they receive if they are aware of which treatments they receive, a serious bias will occur. For example, for the development of topical cream for the indication of some skin disorder, after revelation of the dose, two investigators were asked to give their global evaluation of a patient based on a four-point scale. Despite the fact that the procedures for global evaluation are clearly stated in the protocol, the two investigators gave a rather different evaluation for the patient simply because one of them did not believe that the drug really works at the dose for the patient received and the other one is an advocate for that type of the compound for the treatment of the skin conditions. On the other hand, objective endpoints such as systolic and diastolic blood pressures or total cholesterol levels can be recorded differently if the investigators are aware of treatment assignment. Although some *hard endpoint* such as survival (mortality) or incidence of myocardial infarction are more objective than other clinical endpoints, these can still be subjective. For example, the determination of the cause of death or the diagnosis of infarction may be biased if patient's treatment is known. Therefore open-label trials are generally not recommended for comparative clinical trials. In current practice, open-label trials are not accepted as adequate well-controlled clinical trials for providing substantial evidence for

approval by most regulatory agencies such as the FDA, the European Community (EC), and Ministry of Health, Labor, and Welfare (MHLW) of Japan. However, under certain circumstances open-label trials are necessarily conducted. Spilker (1991) provides a list of situations and circumstances in which open-label trials may be conducted. As indicated in Chapter 1, in order to provide some potentially promising medications to the patients with severely debilitating or life-threatening diseases, clinical trials conducted under compassionate plea protocols or treatment IND may be open labeled. In general, open-label trials are less biased if the clinical endpoints are objective outcomes such as overall survival or the incidence of coma.

Ethical consideration is always an important factor, or perhaps the only factor that is used to determine whether a trial should be conducted in an open-label fashion. For example, phase I dose-escalating studies for determination of the maximum tolerable dose of drugs in treating terminally ill cancer patients are usually open labeled. Clinical trials for evaluation of the effectiveness and safety of a new surgical procedure are usually conducted in an open-label fashion because it clearly unethical to conduct a double-blind trial with a concurrent control group in which patients are incised under a general anesthesia to simulate the surgical procedure. Note that premarketing and postmarketing surveillance studies are usually open labeled. The purpose of premarketing surveillance studies is to collect the data of efficacy and safety with respect to the duration of exposure of a broader patient population to the test drug, while the objective of postmarketing surveillance studies is to monitor the safety and tolerability of the drug product.

By definition, a single-blind study is the one in which either the patient or investigator is blind to the assignment of the patient. In practice, a single-blind trial is referred to as a trial in which only the patient is unaware of his or her treatment assignment. As compared with open-label trials, single-blind studies offer a certain degree of control and the assurance of the validity of clinical trials. However, the investigator may bias his or her clinical evaluation by knowing which treatment the patient receives. Spilker (1991) indicates that results of single-blind trials are equivalent to those from open-label trials. Therefore, when a single-blind trial is planned, it is prudent to ask why this trial cannot be conducted in a double-blind fashion.

A double-blind trial is a trial in which neither the patients nor the investigator (study center) are aware of patient's treatment assignment. Note that the *investigator* could mean all of the health care personnel, which include the study center, contract laboratories, and other consulting experts for evaluation of effectiveness and safety of patients in a broader sense. In addition to the patients and the investigator, if all members of clinical project team of the sponsor associated with the study are also blinded, then the clinical trial is said to be triple-blinded. These members include the project clinician, the CRA, the statistician, the programmer, and the data coordinator. In addition to the patient's treatment assignment, the blindness also applies to concealment of the overall results of the trial. In practice, although the project clinician, the CRA, the statistician, the programmer, and the data coordinator usually have access to the individual patient's data, they are generally not aware of the treatment assignment for each patient. In addition the overall treatment results, if any (e.g., interim analyses), will not be made available to the patient, the investigator, the project clinician, the CRA, the statistician, the programmer, and the data coordinator until a decision is made at an appropriate time. A triple-blind study with respect to blindness can provide the highest degree for the validity of a controlled clinical trial. Hence it provides the most conclusive unbiased evidence for the evaluation of the effectiveness and safety of the therapeutic intervention under investigation.

To ensure the success of a triple-blind study, it is recommended that the following be considered:

1. A carefully chosen study design with an appropriate randomization method.
2. A conscientiously selected concurrent control according to study objectives.
3. Adequate conduct of the trial with no apparent protocol violations.
4. Patient compliance.
5. A sufficient power.
6. Appropriate statistical methods for data analysis.

However, the most important factor for the success of a triple-blind study is to maintain the blindness throughout the entire course of the trial by all participants of all three parties. To protect the integrity of blindness, it is helpful to provide in-house training/education to all personnel related to the clinical trial, including those in the analytical laboratory or in pharmaceutical science research and development (R&D). For example, personnel in the department of analytical laboratory are responsible for the assay of blood samples for active ingredients or metabolites for patients in clinical trials, while the pharmaceutical science R&D develops the matching placebos for the clinical trial. Therefore the personnel at analytical laboratory and pharmaceutical science R&D should have a certain understanding of the concept of blindness and its implication for the integrity of clinical trials.

For a clinical trial comparing a new therapeutic agent with a concurrent control, the departments of pharmaceutical science R&D and drug supply/packaging are usually required to manufacture an identically matched control with the same dosage form. A matched placebo should be identical to the active agent in all aspects such as size, color, coating, taste, texture, shape, and odor except that it contains no active ingredient. The study drugs are then packed in an identical container such as a blister pack or a drug kit affixed with a three-part double-blind tear-off label with the study and patient number. Manufacturing of a perfectly matched control requires certain pharmaceutical techniques and packaging skills provided by both departments. Sometimes, however, a perfectly matched control may not be available due to technical difficulties for some doses. In this situation the method of administration should be modified to maintain the blindness. For example, a phase II clinical trial is to be conducted with daily dose of 100 mg, 300 mg, 600 mg, and a placebo to evaluate the dose-response relationship of a drug. To keep the blindness throughout the study, it is necessary to manufacture placebos to match the drug at different doses. Suppose that the department of pharmaceutical science R&D has difficulties in making matched placebos for tablets of 300 mg and 600 mg. However, the manufacturing of matching placebo for the smallest tablets of 100 mg is still possible. In addition, suppose that patients have difficulties in swallowing the largest tablets of 600 mg. In this case we can modify the method of administration based on 100 mg tablets of the active drug and matching placebos as follows to maintain the blindness.

The first arm of 600 mg: Six 100 mg tablets of the new agent.

The second arm of 300 mg: Three 100 mg tablets of the new agent and three 100 mg matched placebo.

The third arm of 100 mg: One 100 mg tablet of the new agent and five 100 mg matched placebo.

Placebo control arm: Six 100 mg matched placebo.

Note that patients should be instructed to take all six tablets at one time (e.g., in the morning) so that the blindness will not be broken due to different time of administration. The above method is known as *multiple-placebo* or *double dummy*. This method is useful when treatments involve two different active agents or two different routes of administration. For example, a clinical trial is conducted to evaluate a once daily sustained-release formulation of an anti-hypertensive agent with its standard three-times-a-day (t.i.d.) immediate release formulation. The matched placebos can be made for each of two formulations. A bottle of the active sustained release formulation (e.g., bottle S) and another bottle of the placebo tablets of the immediate release formulation (e.g., bottle I) are dispensed to the patients assigned to the sustained-release formulation. The patients assigned to the group of immediate-release formulation receive a bottle of placebo tablets of the sustained-release formulation and another bottle of the active immediate-release formulation. Each patient is instructed to take a tablet from bottle S at 8:00 A.M. in the morning and a tablet from bottle I at 8:00 A.M., 2:00 P.M., and 10:00 P.M. In this case the blindness is preserved without matching tablets identically for all formulations and placebos.

Another example is ISIS-2 (1988) in which the treatments are one-hour intravenous infusions of $1.5\ \mu$ of streptokinase and one-month of 160 mg/day enteric coated aspirin. Therefore the corresponding placebo infusion and tablet were manufactured to match the active treatments as described previously in Chapter 2. However, blindness for ISIS-2 is possible because the matched placebo infusion has the same one-hour IV infusion at the same rate. On the other hand, the arm of accelerated rt-PA in the GUSTO I trial (1993) had a bolus dose of 15 mg, 0.75 mg per kg of body weight, over a 30-minute period, not to exceed 50 mg; and 0.5 mg per kg, up to 35 mg, over the next 60 minutes. The active control arms used the same one-hour infusion of $1.5\ \mu$ of streptokinase as ISIS-2. Therefore, because the dose of the accelerated rt-PA had to be adjusted for body weight twice during the infusion, IV heparin had to be titrated according to the activated partial-thromboplastin time and its length of infusion was also different from other arms receiving streptokinase. As a result the GUSTO I study was an open-label study. Although primary efficacy outcome is the mortality from stroke and bleeding complication as the primary safety endpoint. However, they are subjective to possible bias if the treatments are known to investigators, in particular, when classification of stroke and bleeding requires clinical judgment for some borderline cases. Due to the large size of the GUSTO I study, the bias could be accumulated rapidly and become serious just from some subtle, consciously, sub-consciously, or unconsciously error in clinical judgment made by an investigator. The GUSTO I study, however, failed to address the bias issue due to the open label. As a result there were tremendous debates over the fact that the GUSTO I was an open-label study (Rapaport, 1993; Sleight, 1993; Rider et al., 1993). In their response to the rebuttal article by the investigators of the GUSTO I trial (Rider et al., 1994; Lee, 1994), Rider et al. (1994) state the essence of randomization and blindness in clinical trials: "randomization of patients is done to try to ensure that no major differences exist in baseline characteristics between treatment groups before treatments are administered, double-blinding is done to ensure that no differential effects occur after treatments are given." It is sad to see that the breach of blindness, the omitting of a rather routine and operationally and economically feasible insertion of an extra intravenous line, casts a serious shadow over the scientific validity of this originally spotless trial, and introduces an inadvertent and impossibly assessed bias.

For multiple placebos the so-called method of the *multiple-evaluator* is useful to preserve the blindness. For example, suppose that a clinical trial is conducted in three doses to

assess the dose-response relationship of a contrast-enhanced agent in conjunction with magnetic resonance imaging for the diagnosis of malignant liver tumors in patients with known focal liver lesions. Since the contrast agent is administered as an IV injection by reconstruction from the vial of an active agent and the vial of a saline solution according to the body weight of each patient, the clinician who prepares and administers the contrast medium will know the dose. If the clinician is also responsible for the evaluation of the results of pre- and postcontrast MRI, bias will occur during the evaluation of films and safety data due to prior knowledge of the doses. In this case the multiple-evaluator method is helpful. At each study center, one clinician will prepare the injection according to the randomization codes in total privacy without divulging the dosing information to anyone in the study center. The clinician will then administer the contrast medium without showing the syringe to everyone. The other clinicians at the study center will evaluate the films in a totally blinded manner. The multiple-evaluator method is also useful in physical therapy. A clinical trial was conducted to evaluate the transcutaneous electrical nerve stimulation (TENS) for patients with chronic low back pain (Deyo et al., 1990). This trial employed a two-group parallel design with a real TENS and a sham TENS group. Although it is known to be very difficult to implement, in order to maintain blindness over the entire course of the study, the therapist who is responsible for instructing patients and applying TENS, asked patients not to discuss their therapy with the clinician who performed the evaluations. The clinician who performed the evaluation at baseline is different than the one at the follow-up visit. In addition the frequency of visits and duration of treatment were identical for the two groups, as were all written and verbal instructions and effort to identify ideal electrode placement.

In practice, even with the best intentions for preserving blindness throughout a study, blindness can sometimes be breached for such reasons as a distinct adverse event or the taste of the active treatment. One method to determine whether the blindness is seriously violated is to ask both patients and investigators to guess the patient's treatment assignment during the study or at the conclusion of the trial prior to unblinding. Once the guesses by patients and investigators are recorded on the case report forms and entered into the database, the degree of unblinding and its impact on introducing bias in the evaluation of treatment effect can be assessed. In what follows, some examples that may be of interest for practical use are adopted from the literature.

For example, a one-year double-blind placebo-controlled study was conducted by the NIH to evaluate and distinguish between the prophylactic and therapeutic effects of ascorbic acid for the common cold (Karlowksi, 1975). 311 employees of NIH were randomly assigned to receive the active agent or the matched placebo based on the method of complete randomization. One hundred and ninety of them completed the study. In this study, since there was no time to design, test, and manufacture a perfectly matched placebo for ascorbic acid due to the seasonal constraint. At an early stage of the study the researchers discovered that some subjects had tasted the contents of their capsules and professed to know which treatment they were taking. At the completion of the study, in order to assess the bias, a questionnaire was distributed to everyone enrolled in the study so that they could guess which treatment they had been taking. Table 4.6.1 presents the results from the 190 completed subjects for the prophylactic use. The number of correct guesses was 79 and the number of misses 23. Therefore the expected bias factor is estimated to be 28 (a half of the difference between 79 and 23). Hence considerable selection bias occurred in this study. Note that the association between the severity and the duration of symptoms and knowledge of the medication taken were also established by the researchers of this project.

Table 4.6.1 Results of Patient's Guess on Treatment for the Prophylactic Use

Patient's Guess	Actual Assignment	
	Ascorbic Acid	Placebo
Ascorbic acid	40	11
Placebo	12	39
Do not know	49	39
Total	101	89

Source: Karlowski et al. (1975).

Note that to test the integrity of blinding, Chow and Shao (2003) propose a method of testing treatment effects by incorporating the data of patients' guesses of their treatment codes. The idea is to include the patients' guesses as a factor in the analysis of variance (ANOVA) for the treatment effects.

The entire process of a clinical trial involves many activities by many personnel. Blinding should be applied to all participants performing activities/functions at every stage of the entire course of a clinical trial. However, for some occasions during a clinical trial, unblinding the treatment codes for individual subjects may be necessary. For example, the occurrence of serious adverse events may necessitate the breaking of the treatment code of the patients. If the sponsor's staff, including bioanalytical scientists, auditors, and those responsible for serious adverse event reporting who are not involved in the treatment or clinical evaluation of the subjects, are required to be unblinded to the treatment codes, the ICH E10 guidance suggests that the sponsor develops adequate standard operating procedures (SOPs) to guard against inappropriate dissemination of the treatment codes. However, as indicated by the ICH E10 guidance, breaking the blind should only be considered when is deemed essential by the patient's primary care physician. In addition, any intentional or unintentional breaking the blind should be reported and explained irrespective of the reason for its occurrence. The procedure and timing for revealing the treatment assignments should be adequately documented. Furthermore, checking, editing, and evaluation of data and preparation of statistical analysis plan (SAP) should also be conducted in a blinded fashion and should be finalized before the database for the trial is locked and treatment codes are unblinded. Proper documents should chronically record the locking of the database, the unblinding of the treatment codes of individual subjects, and statistical analysis according to standard operating procedures.

In practice, similar issues for the maintenance of blindness are commonly seen in other therapeutic areas. For example, beta-blocker agents (e.g., pro-pranolol) have specific pharmacologic effects such as lowering blood pressure and the heart rate and distinct adverse effects such as fatigue, nightmares, and depression. Since blood pressure and heart rate are vital signs routinely evaluated at every visit in clinical trials, if a drug such as propranolol is known to lower blood pressure and the heart rate, then preservation of blindness is a huge challenge and seems almost impossible. The Beta-Blocker Heart Attack Trial (BHAT) is a landmark, multicenter, double-blind, randomized, placebo-controlled trial designed to test the effectiveness of beta-blocker in reducing mortality during a two- to four-year period in postmyocardial infarction patients (Beta-Blocker Heart Attack Trial Research Group, 1982, 1983). At the conclusion of the trial, patients, investigators, and clinic coordinators were asked to guess the patient's treatment assignment. Table 4.6.2 provides the proportion of correct guesses and estimates of the expected bias factor. Apparently blindness was not totally

Table 4.6.2 Proportions (%) or Correct Guesses for Beta-Blocker Heart Attack Trial

	Propranolol	Placebo	Estimate of the Expected Bias Factor
Patient	79.9	42.8	380 ($N = 3230$)
Investigator	69.6	68.6	568 ($N = 3398$)
Clinic coordinator	67.1	70.6	669 ($N = 3552$)

maintained even for this landmark study with a major influence on management of care for patients who suffer myocardial infarction. Morgan (1985) suggested that to quantify possible bias, the researchers for BHAT should administer the questionnaire of guesses of patient's treatment three months into the trial rather than at the end.

4.7 DISCUSSION

ICH E10 guideline points out that the objective of using a control group in a clinical trial is to allow discrimination of subject outcomes caused by the test treatment from outcomes caused by other factors. In order to achieve this goal, subjects in the test treatment and control groups should be as similar as possible with respect to all baseline and on-treatment variables, such as the design, conduct, analysis, and interpretation of the results, that could influence outcome other than the study treatment. Failure to achieve this similarity may introduce bias. Randomization and blinding are the two techniques to guarantee this similarity. Randomization is the technique to ensure that the test treatment and the control groups are similar at the beginning of the trial. On the other hand, blinding is employed to make sure that the two groups are treated similarly during the course of the trial. As a result, control, randomization, and blinding are three key features for a critical determinant of quality and persuasiveness of a clinical trial.

As blinding is a key to the integrity of randomized controlled trials, Devereaux et al. (2001) conducted a survey among attending physicians and textbook on blinding terminology. Ninety-one attending physicians at Dalhousie University, McMaster University, and University of Calgary Foothills Hospital completed a survey that defined the six groups, who are potential candidates for blinding in a randomized, controlled trial. These six groups include subjects, health care providers, data collectors, data analysts, judicial assessors of the outcomes, and personnel writing the paper or report. Responders offer their opinions on which group should be single-, double-, or triple-blinded. They reported that physician respondents identified 10, 17, and 15 unique interpretations of single-, double-, and triple blinding, respectively. They also surveyed 25 textbooks published since 1990 on the definition of blinding using the terms clinical epidemiology, randomized controlled trials, and evidence-based medicine. These 25 textbooks provide five, nine, and seven different definitions of single, double, and triple blinding, respectively. In addition, since June 2000, they also surveyed 200 recently randomized, controlled trials published in the *Annals of Internal Medicine*, *BMJ*, *JAMA*, *The Lancet*, and *The New England Journal of Medicine*. Among these 200 randomized, controlled trials, 5 are single-blinded and 83 are double-blinded. In the 5 single-blind trials, 1 trial did not mention which of the 6 groups was blinded, 2 trials identified 1 group that was blinded, and 1 trial identified 2 groups that were blinded. For the 83 double-blind studies, 41 did not mention which of the six groups were blinded, 29 trials identified 1 group that was blinded, 11 trials identified 2 groups that were blinded, 1 trial

identified 3 groups that were blinded, and 1 study identified 4 groups that were blinded. Therefore, there is a great variation in the interpretation and definition regarding the concept of single, double, and triple blinding by both physicians and textbooks. Devereaux et al. (2001) suggest that explicit statements about the blinding status of specific personnel involved in randomized, controlled trials be reported for the paper published by medical journals. It is interesting to know that the 25 textbooks did not include some well-known books on clinical trials, including the first edition of this book (Chow and Liu, 1998; Piantadosi, 1997; Pocock, 1996; Wooding, 1994; Friedman et al., 1998; Gilbert, 1992).

Ioannidis et al. (2001) conducted a meta-analysis to compare evidence of treatment effects between randomized and nonrandomized trials. They searched MEDLINE (1966– March 2000) and Cochrane Library (Issue 3, 2000) for candidate trials. Two hundred and forty randomized clinical trials and 168 nonrandomized trials in 45 diverse topics with binary outcomes as one of primary measures for efficacy were included in the meta-analysis. They reported that a very good correlation between the summary odds ratios of randomized and nonrandomized trials ($r = 0.75, p < 0.001$). However, nonrandomized trials tended to present large treatment effects. In addition, among the 45 topics, the natural logarithm of the odds ratios differed by at least 50% in 62% of the 45 topics. In addition, in 33% of the 45 topics, the odd ratios vary by at least two-fold between nonrandomized trials and randomized studies. Between-study heterogeneity of treatment effects was also more frequently observed in nonrandomized studies than that among randomized trials. Despite a good correlation between randomized and nonrandomized trials, the treatment effect observed in nonrandomized studies may be overexaggerated and is more heterogeneous among trials.

As indicated earlier, randomization is integral to the success of clinical trials that address scientific and/or medical questions. However, it should be noted that in many clinical research situations, randomization may not be feasible. For example, nonrandomized observational or case-controlled studies are often conducted to study the relationship between smoking and cancer. As a result, in the report entitled *Smoking and Health* by the U.S. Surgeon General issued in 1964, seven key nonrandomized observational studies were cited as the evidence for the relationship between smoking and cancer. Note that if the randomization is not used for some medical considerations, the FDA requires that statistical justification be provided with respect to how systematic selection bias can be avoided.

It should be noted that in practice, for most clinical trials patients are enrolled into study sites in a nonrandom fashion. The selection of study sites is also a nonrandom process. Consequently the validity of statistical inference on the targeted patient population is seriously in doubt. It is therefore recommended that appropriate statistical methods be derived based on the method of the selected randomization model.

Further, although triple blinding is reserved for large cooperative, multicenter studies monitored by a committee, it has nevertheless been applied to company monitors to ensure that they remain unaware of the treatment allocation. In general practice, double blinding is the standard, since it provides the greatest probability for reducing bias.

Blocking is usually employed to ensure that the number of patients in each treatment group will be similar at certain points. For this purpose, small block sizes such as 2, 4, 6, and 8 are usually chosen. Within each block, patients are randomly allocated to receive either the treatment or a control. In some situations, however, deliberate unequal allocation of patients between treatment groups may be desirable. For example, it may be of interest to allocate patients to the treatment and the control in a ratio of 2 to 1. This is a consideration in situations where (1) the patient population is small, (2) previous experience with the study medicine is limited, and (3) the response profile of the competitor is well known.

5

DESIGNS FOR CLINICAL TRIALS

5.1 INTRODUCTION

As was discussed in Chapter 3, the first step in selecting an appropriate statistical design is to determine the objective(s) of the proposed clinical trials. The objective(s) of a clinical trial is usually to answer one or more scientific or medical questions related to the therapeutical intervention under study. Once the study objective(s) have been carefully defined, an appropriate (or optimal) design for the intended clinical can be chosen. Since a wrong choice of design may result in a worthless study, a good statistical design is regarded as an essential prerequisite of clinical trials. Spilker (1991) indicates that choosing the most appropriate design for a clinical trial is similar to choosing ready-made clothes. Temple (1982) indicates that the selection of the most appropriate design or the optimal design depends on the questions asked. The questions that must be asked before choosing an appropriate design include the study objective(s), the nature of the study drug, the disease status/condition under investigation, and other considerations as described in previous two chapters. Therefore, the FDA suggests that a statement of the specific objectives of the study be provided. To clarify the study objective(s), the following questions are helpful.

1. What aspects are being studied?
2. Is it important to investigate other issues that may have an impact on the study drug?
3. Which control(s) might be used?

Once the objective(s) of the study is clearly stated, it is important to determine the aspects that will be studied. These aspects include the dosage form, dose, and the intended indication. For

an indication under investigation, an appropriate dosage form is necessary chosen for the targeted patient population so that the drug can be delivered to the site of action efficiently for optimal therapeutic effect. In addition the selected dose may have an impact on the assessment of the effectiveness and safety of the study drug. For example, a low dose may show a better safety profile, and yet it may not produce clinically meaningful efficacy of the study drug. On the other hand, a high dose may cause a serious side effect. In some clinical trials, the dose may be required to be titrated during the study in order to reach the optimal therapeutic effect. For dose titration, the titration procedures must be clearly described in the protocol. The commonly employed titration procedures include forced titration and titration based on clinical outcome whereby the dose is titrated upward at intervals until intolerance or some specific endpoint is achieved. The issues of possible drug-to-drug interactions with food and/or other concomitant medications, the impact of patient compliance, and pharmaco-economic outcomes such as quality of life associated with the efficacy and safety of the study drug should also be considered in choosing an appropriate statistical design. In addition, as was indicated in Chapter 3, it is also important to determine what control(s) will be used for comparative clinical trials. Different controls may serve different purposes of a clinical trial.

Selecting an appropriate statistical design is critical in clinical development during the process of drug development. In practice, when a new test agent reaches the stage for clinical development, its pharmacological/pharmacokinetic properties and the effectiveness and safety may have been studied through *in vitro* laboratory testing and *in vivo* animal studies. At this time point, however, the safety and effectiveness in humans are not known, and the test agent must be rigorously and scientifically evaluated through clinical trials within the confinement of regulations. As indicated in Chapter 1, the purposes of phase I and early phase II studies are not only to characterize the safety profile but also to determine the therapeutic range of the test agent. Since the test agent is never tested in humans, it is a challenge to acquire information regarding early safety and efficacy of the test agent. The information is extremely helpful for planning of subsequent trials. To capture as much needed information as possible, the utilization of an efficient statistical design is critical.

In recent years, there has been tremendous discussion on whether the choice of study design should be based solely on medical consideration. Another interesting question raised is whether to include marketing, regulatory, and/or statistical perspectives as well. Ideally an optimal design will account for considerations from different perspectives. In practice, however, such a design may not exist. It should be noted that considerations from different perspectives always mean limitations to the choice of design. Therefore, Temple (1982) points out that a study must be sufficient to its task, and design limitations should be understood before proceeding, first to see whether a better design can be found and to understand the limits on interpretation imposed by a less than optimal design, and second, so that, if necessary, the limits can be discussed with the regulatory agency and potential problems anticipated.

When planning a clinical trial, it is suggested that the relative merits and disadvantages of candidate statistical designs be compared before an appropriate design is chosen for the clinical trial. It is important to evaluate the suitability of the chosen design for addressing scientific/medical questions and/or claims. For example, if we are to choose between a crossover design and a parallel design for a clinical trial, we must first understand the nature of these two designs. For a parallel design, each patient receives one and only one treatment in random fashion, whereas for a crossover design each patient receives more than one treatment at different dosing periods. If a clinical trial is intending to investigate the residual effect that may be carried over from one treatment to another, a crossover design could be employed. Note that the Federal Register (Vol. 42, No. 5, Sec. 320.26(b) and 320.27(b), 1977) indicate

that a bioequivalence trial (single dose or multiple dose) should be crossover in design, unless a parallel design or another design is more appropriate for some valid scientific reasons. On the other hand, if a clinical trial is intended to demonstrate the effectiveness and safety of a study medicine, a parallel design is more appropriate.

In the next section, we introduce parallel designs, including parallel group designs and matched pairs parallel designs. Section 5.3 describes clustered randomized designs that have been extensively employed in community-based intervention clinical trials. This section also addresses the difference between an experimental unit (a unit of randomization) and a unit of analysis. Section 5.4 discusses several different types of crossover designs. Section 5.5 covers titration designs, including some variations such as a forced dose-escalation design. The concept of enrichment designs is given in Section 5.6. Sections 5.7, 5.8, and 5.9 examine group sequential designs, placebo challenging designs, and blinder reader designs, respectively. In the last section, we provide a discussion regarding the selection of an appropriate design.

5.2 PARALLEL GROUP DESIGNS

A parallel group design is a complete randomized design in which each patient receives one and only one treatment in a random fashion. Basically there are two types of parallel group design for comparative clinical trials, namely, group comparison (or parallel-group) designs and matched pairs parallel designs. The simplest group comparison parallel group design is the two-group parallel design which compares two treatments (e.g., a treatment group vs. a control group). Each treatment group usually, but not necessarily, contains approximately the same number of patients. The ICH E9 guideline “*Statistical Principles for Clinical Trials*” indicates that the parallel group design is the most common trial design for confirmatory trials (ICH E9, 1998). An example of a three-group parallel design with a test treatment and two controls (e.g., an active control A and a placebo control B) is illustrated in Figure 5.2.1. This is a typical example of multiple controls (ICH E10, 1999). The use of this three-group parallel design with an active control and a placebo control can distinguish an ineffective drug from an ineffective design. An effective design can be determined based on the evaluation of *assay sensitivity* by showing the superiority of the active control over placebo. As indicated by ICH E10 guideline, this design is particularly useful when the test drug and the placebo provide similar results. This is because it provides

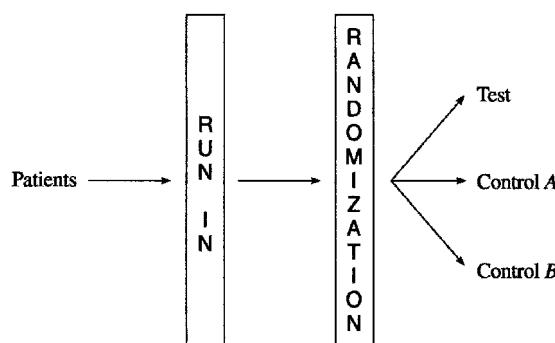


Figure 5.2.1 Parallel group design.

evidence that the test drug has little or no efficacy as compared to the placebo. On the other hand, if neither the test drug nor the active control can be distinguished from the placebo in terms of efficacy, this clinical trial is said to be lack of assay sensitivity, and hence, it does not provide evidence to conclude that the test drug is effective. Some of major advantages of a parallel group design are summarized below:

1. It is simple and easy to implement.
2. It is universally accepted.
3. It is applicable to acute conditions (e.g., infection or myocardial infarction).
4. Analysis is less complicated, and interpretation of the results is straightforward.

In addition, for ethical consideration with the control (e.g., the placebo), we can allocate patients unequally to treatment groups (in a random fashion) to allow more patients to receive the treatment (e.g., in a 2 to 1 or 3 to 1 ratio). A parallel group design is probably the most commonly used design in phases II and III of clinical trials. However, it usually requires more patients than other comparative designs.

The matched pairs parallel group design is a randomized complete block design with a block size of 2 in which each patient is *matched* with another of similar prognostic characteristics (e.g., obesity) for the disease under investigation. One patient in each pair is assigned the treatment, and the other receives the control. As compared to parallel group designs, matched pairs parallel group designs can reduce variability from treatment comparison. In addition a matched pairs parallel group design requires a smaller patient population. Therefore it is considered a more suitable design for progressive diseases such as cancer. However, matched pairs group designs suffer the disadvantages such that (1) the prognostic characteristics are not easily defined and (2) patient recruitment is usually slow. In practice, the selection bias for matched pairs designs is usually a concern in patient recruitment, which often limits its applications in clinical trials. Note that a matched pairs design is in fact an extreme case of stratification which is often considered to achieve balance in covariates or prognostic factors. When the number of covariates is large, the matched pairs design is difficult to implement. Hence the matched pairs design is not of practical interest in this case. Although at the planning stage it is almost impossible to identify all of the covariates that may have an impact on the disease, an unbiased estimate of the treatment effect can still be obtained by adjusting these covariates in analysis regardless of whether they are used for stratification or matching in order to achieve the balance in covariates.

In clinical trials, for a given clinical endpoint, basically there are two kinds of variability associated with the response. These two kinds of variability are known as the interpatient and intrapatient variabilities. Statistically the smaller these variabilities are, the more accurate and reliable the clinical results will be. For a parallel group design, however, these variabilities cannot be identified because each patient receives the same treatment during the entire course of the study. In other words, the observed variability for any comparisons between groups contains both interpatient and intrapatient variabilities that cannot be separated and estimated due to the nature of the parallel group design. As a result a parallel design does not provide independent estimates of the interpatient and intrapatient variabilities. In practice, a parallel group design is an appropriate design for comparative clinical trials if the interpatient variability is relatively small compared to the intrapatient variability. This is because a valid and efficient comparison between treatment is often assessed based on the intrapatient variability. Therefore, if the interpatient variability is relatively small

compared to the intrapatient variability, the observed variability will be close to the intrapatient variability. In this case the parallel group design will provide a more accurate and precise assessment of the treatment difference. Other considerations for the use of parallel group designs include patient characteristics (e.g., acute or chronic and very ill or life threatening) and the nature of study medicine (e.g., potential toxicity and long elimination half-life). In some cases financial consideration may be a key factor for selecting parallel designs.

Run-in Periods

Before patients enter a clinical trial, a run-in (or lead-in) period of placebo, no active treatment, dietary control, or active maintenance therapy (e.g., diuretic and/or digoxin in heart failure studies) is usually employed prior to randomization. The inclusion of a run-in period prior to the active treatment has the following advantages:

1. It acts as a washout period to remove effects of previous therapy.
2. It can be used to obtain baseline data and to evaluate if patient fulfills study entry criteria.
3. It can be used as a training period for patients, investigators, and their staff.
4. It helps in identifying placebo responders.
5. It provides useful information regarding patient compliance.
6. It can be used to estimate and compare the magnitude of possible placebo effects between groups.

In clinical trials it is desired to have a washout period prior to an active treatment period to wear off effects of previous therapy for an unbiased and valid assessment of the study medicine. A run-in period, however, may not be suitable for patients whose conditions are acute requiring immediate treatment. It is acceptable if patients can remain without active therapy for a short period of time. In many clinical trials, it is not uncommon to observe the placebo effect for many drug products. For example, for antidepressant agents, an intensive care period may significantly improve the patients' depression without any treatment. At the end of active treatment period, it is important to determine whether the observed significant effect is due to the placebo or treatment. To eliminate the possible placebo effect, it is suggested that a run-in period be included to establish patient comparability between treatment groups at baseline, and this helps to remove placebo effect from comparison at the endpoint evaluation. In clinical trials, patients' cooperation and/or their compliance to study medicine is always a concern. A run-in period can be used as a training period for patients, investigators, and their staff. For example, if the trial requires patients to complete diary cards, a run-in period provides a training period for the patients to be familiar with the diary cards. In addition it may help in identifying uncooperative patients at an early stage and provide the necessary counseling. This information is useful in improving a patient's compliance when the study moves to the active treatment period.

Note that a run-in period is usually employed based on a single-blind fashion. In other words, the participated patients are not aware of receiving a placebo. Although the inclusion of a run-in period in clinical trials has many advantages, it increases the length of a study; consequently it often requires extra study visits. This has a direct impact on the increase of cost and potentially a decrease in enthusiasm by patients and investigators.

Examples of Parallel Group Design in Clinical Trials

During clinical development of a drug product, parallel group designs are often considered to evaluate the efficacy and safety of a monotherapy or combination therapy with other agents of the drug product. In addition a parallel group design may be used to study the dose response of a drug product. For example, consider the clinical development of *Glucophage* (metformin hydrochloride). Glucophage is an oral agent for the treatment of type II noninsulin-dependent diabetes mellitus (NIDDM). Although Glucophage has been on the European market for more than 20 years, it was not available for the U.S. market until it was approved by the FDA in the late December 1994. Over the past few years a number of clinical trials were conducted to further investigate the clinical pharmacology and other uses of the drug product.

To illustrate the application of parallel group designs in clinical trials, consider the clinical development of Glucophage. Table 5.2.1 lists three studies of Glucophage regarding evaluation by monotherapy, combination therapy with insulin, and dose response. For the first study (Dornan et al., 1991), the objective was to test the efficacy and tolerability of Glucophage. This study was an eight-month double-blind placebo-controlled trial of Glucophage monotherapy in 60 obese patients with NIDDM. This study had a typical parallel group design with a run-in period. After a dietetic review and a one-month run-in period, patients were stratified according to the levels of glycosylated hemoglobin (H_bA_{1C}) concentration and randomized to receive either Glucophage or an identical dose of placebo. The starting dose was one tablet (500 mg) daily increased at weekly intervals to three tablets daily after one month. Thereafter the dose was increased by one tablet daily at weekly intervals to a maximum of two tablets three times daily, aiming for lowering the level of fasting blood glucose less than 7 mM (7 mmol per liter or 126 mg per deciliter, mg/dL). Patients were fasted at the beginning and end of the run-in period, and after 1, 3, 5, and 8 months of treatment they were weighted and their blood pressure was measured. In addition, blood was taken for fasting glucose, total cholesterol, triglycerides, H_bA_{1C} , and serum insulin. The results indicated that Glucophage reduced H_bA_{1C} levels from 11.7% to 10.3%, whereas the placebo treatment resulted in a rise from 11.8% to 13.3%. The mean percent reduction in H_bA_{1C} of Glucophage is 23% lower than the placebo without weight gain. In addition the final mean fasting blood glucose level was 5.1 mM (92 mg/dL) lower on Glucophage than on the placebo. The fasting glucose level fell from 13.5 (243 mg/dL) to 10.2 mM (184 mg/dL) (about 24%) on Glucophage and rose from 12.7 (229 mg/dL) to 15.3 mM (275 mg/dL) (about 17%) on the diet plus placebo. No changes or differences between groups were observed in body weight, blood pressure, C peptide,

Table 5.2.1 Examples of Parallel-Group Design

Study	Purpose	Sample Size	Parallel Groups	Duration (Run-in + Active)	Primary End Point
Dornan et al. (1991)	Monotherapy	60	2	1 mo + 8 mo	H_bA_{1C} , PG
Giugliano et al. (1993)	Combination therapy	50	2	4 wk + 6 mo	H_bA_{1C} , FPG
Bristol-Myers Squibb (1994)	Dose response	360	6	3 wk + 11 wk	H_bA_{1C} , FPG

Note: PG = plasma glucose; FPG = fasting plasma glucose; H_bA_{1C} = hemoglobin A_{1C} .

serum insulin, or triglycerides. As a result Dornan et al. (1991) concluded that Glucophage monotherapy is an effective and well-tolerated first-line treatment for obese patients with NIDDM. They also indicated that the use of Glucophage should not be restricted to very obese patients because Glucophage lowers HbA_{1C} and achieves approximately equivalent improvements in glycemic control in both mildly and moderately to severely obese patients.

Another application for the use of a parallel group design would be the evaluation of combination therapy of the current insulin regimen with Glucophage. Giugliano et al. (1993) studied the efficacy and safety of Glucophage in the treatment of obese NIDDM patients poorly controlled by insulin after secondary failure to respond to sulfonylurea. The study is a typical parallel group design consisting of a four-week run-in single-blind phase and a six-month double-blind treatment phase. During the placebo run-in phase, patients were given the current insulin regimen and asked to maintain their regular diet. After a six-month active treatment, Glucophage was shown to have significantly improved the glycemic and lipid control. The results indicated that after four months, the glucose level declined by 31% (4.1 mM or 73.8 mg/dL) from baseline, H_bA_{1C} levels by 1.7%, and fasting insulin levels by 26%. In addition the necessary insulin dose was also reduced by more than 20% (from 90 to 71 U/d). Furthermore in the Glucophage group there were significant changes from both the baseline and placebo in levels of total cholesterol (-0.21 mM), triglycerides (-0.31 mM), and high-density lipoprotein cholesterol (+0.13 mM), and blood pressure was reduced an average of 8.8 and 4.8 mmHg versus the baseline and placebo, respectively. Therefore, Giugliano et al. (1993) concluded that combination Glucophage therapy represents a safe and efficacious strategy for improving glycemic regulation and coronary artery disease risk status in patients with NIDDM which was inadequately controlled by insulin alone.

Recently, to fulfill the FDA's requirement, a study was conducted by Bristol-Myers Squibb to study the dose response of various dose levels of Glucophage compared to a placebo in patients with NIDDM. The study was a randomized double-blind placebo-controlled parallel-group study that consisted of two phases. At the end of single-blind placebo run-in period, qualified patients were randomized to one of the six double-blind treatment groups (i.e., placebo, Glucophage at 500 mg, 1000 mg, 1500 mg, 2000 mg, and 2500 mg per day) for 11 weeks. Patients assigned to the dose levels of Glucophage 500 mg/d or Glucophage 1000 mg/d began the active treatment phase at this dose level and continued on it throughout the study. Patients assigned to dose levels of Glucophage 1500 mg/d, 2000 mg/d, or 2500 mg/d underwent a forced titration during the initial three weeks of study to minimize the possibility of gastrointestinal side effects. All patients were maintained for a minimum of eight weeks on their final assigned dose level. The results suggest that there is a dose response showing Glucophage to be effective at all randomized dose levels. The dose response increased up to the 2000 mg dose but then decreased as the dose was increased from 2000 mg to 2500 mg. These results are consistent with those for FPG and H_bA_{1C} at treatment weeks 7, 11, and the end of the trial. Given the dose levels considered in this study, it is concluded that 500 mg is the minimum effective dosage ($p = 0.03$) and 2000 mg is the maximum effective dose level ($p = 0.001$) compared with the placebo.

Another example concerning the study of metformin with a parallel group design is the Diabetes Prevention Program (1999, 2002). As risk factors associated with type 2 diabetes, such as elevated plasma glucose concentrations in the fasting state and after an oral glucose load, overweight, and sedentary lifestyle are potentially reversible, the Diabetes Prevention Program Research Group hypothesized that modifying these factors with

a lifestyle-intervention program or the administration of metformin would prevent or delay the development of diabetes. Therefore, the following three primary scientific questions were of particular interest to the Diabetes Prevention Program Research Group, which were intended to be answered by this study:

1. Does a life intervention or treatment with metformin, a biguanide antihyperglycemic agent, prevent or delay the onset of diabetes?
2. Do these two interventions differ in effectiveness?
3. Does their effectiveness differ according to age, sex, or race or ethnic group?

The Diabetes Prevention Program Research Group employed a randomized, placebo-controlled three-group parallel design in order to address these three questions. A total of 3,234 nondiabetic subjects with elevated fasting and post-load plasma glucose concentrations were randomized to one of three interventions: standard lifestyle recommendations plus metformin at a dose of 850 mg b.i.d., standard lifestyle recommendations plus placebo b.i.d. or an intensive program of lifestyle modification with the goals of at least a 7% weight loss and at least 150 minutes of physical activity per week. This study initially included a fourth treatment group, i.e., troglitazone at a dose of 400 mg q.d., which was discontinued in 1998 due to its potential liver toxicity. Except for the group of a lifestyle modification program, a double-blinded technique was applied in this study to minimize bias.

5.3 CLUSTER RANDOMIZED DESIGNS

The fundamental theory of the classic experimental design by Fisher (1947) is based on the fact that the randomization unit is the same as the analysis unit used as the experimental unit for statistical inference. Statistical inference based on the principle of randomization unit being the analysis unit is hence the most efficient in the sense that it produces the maximum power and requires the minimum sample size. Almost all clinical trials for evaluation of therapeutic intervention have adopted such a principle. For example, subjects such as patients or normal volunteers mentioned throughout this book are not only the unit of randomization but also a unit of statistical inference. However, on the other hand, for assessment of nontherapeutic interventions such as lifestyle intervention or new educational program for smoking cessation, randomization may be easily performed and trials can be efficiently implemented to reduce bias through randomization of some social intact units such as family, school, worksites, athletic teams, hospitals, or communities. These intact social units are called *clusters*. This type of design is hence referred to as *cluster randomized design* or group randomized design (Donner and Klar, 2000).

For cluster randomized designs, randomization is performed at the cluster level rather than at the subject level. Thus, the unit of analysis may not be necessarily the same as the unit of randomization. If the inference is made at cluster level, then the standard methodologies for traditional clinical trials provided in this book can be applied because cluster is the unit of randomization as well as the unit of analysis. However, for most clinical trials with a cluster randomized design, the inferences are intended at the subject level, and hence, the standard methods for sample size calculation and data analysis considering subject as analysis unit are not appropriate. One of the major considerations for design and analysis of cluster randomized trials is the control of the intracluster and intercluster variations. As clusters are some intact social units such as families or worksites, therefore, we would anticipate that the subjects

within the same cluster might share the same traits or have similar characteristics. In other words, the subjects within the same cluster are more similar than are those between clusters. One statistical measure to quantify this similarity is the intraclass correlation coefficient (ICC) (Snedecor and Cochran, 1980). If the intraclass correlation coefficient, denoted by ρ , is positive, the intracluster variation is smaller than the intercluster variation. The ICC plays a very important role in analysis of cluster randomized trials using subjects as the unit of inference.

Denote by Y_i some clinical endpoint for subject i from a random sample of n subjects. Suppose that Y_i is normally distributed with population mean μ and variance σ^2 . Let \bar{Y} be the sample mean of the endpoint from these n subjects. The variance of the sample mean is given by

$$V(\bar{Y}) = \sigma^2/n.$$

Suppose that instead of a single random sample, there are k clusters of m subjects each for a total of n subjects, i.e., $n = km$. It can be easily verified that under the assumption of a constant ICC for all k clusters, the variance of the sample mean is given by

$$\begin{aligned} V(\bar{Y}) &= (\sigma^2/km)*[1 + (m - 1)\rho] \\ &= \sigma^2/n*\text{VIF} \end{aligned} \quad (5.3.1)$$

where VIF denotes the variance inflation factor. It can be seen from (5.3.1) that the variance of the sample mean under cluster randomized design increases by the variance inflation factor, which is a function of both ICC and cluster size m . Table 5.3.1 gives the variance inflation factors for ICC = 0.02 to 0.2 by 0.02 and m = 20 to 100 by 20. Table 5.3.1 reveals that the variance of the sample means under the cluster randomized design inflates very rapidly as the ICC and cluster size increase. Even for the situation that ICC is as negligible as 0.02 and the cluster size is as small as 20, the variance of the mean under the cluster randomized design is still about 38% greater than that under the traditional randomization by individual subjects. In addition, the sample size required for the cluster randomized design is the multiple of the sample size needed for randomization at individual subject level. For example, when ICC = 0.06 and the cluster size is 100, then the trials

Table 5.3.1 Variance Inflation Factors for Various Combination of Intraclass Correlation Coefficient and Cluster Size

ICC	Cluster Size				
	20	40	60	80	100
0.02	1.38	1.78	2.18	2.58	2.98
0.04	1.76	2.56	3.36	4.16	4.96
0.06	2.14	3.34	4.54	5.74	6.94
0.08	2.52	4.12	5.72	7.32	8.92
0.10	2.90	4.90	6.90	8.90	10.90
0.12	3.28	5.68	8.08	10.48	12.88
0.14	3.66	6.46	9.26	12.06	14.86
0.16	4.04	7.24	10.44	13.64	16.84
0.18	4.42	8.02	11.62	15.22	18.82
0.20	4.80	8.80	12.80	16.80	20.80

with a clustered randomized design require a sample size almost seven times as large as that of the clinical trials using the traditional individual randomization.

Two commonly encountered mistakes in most cluster randomized trials are given below:

1. Although the trials adopt a cluster randomization, the analysis of data completely ignores this fact and uses subject as the unit of analysis.
2. Sample size estimation fails to take into consideration the variance inflation factor.

The first mistake ignores the intercluster variation and hence could not control the false-positive rate at the prespecified significance level. The second mistake results in an underestimation of the required sample size and hence may increase the false-negative rate. Cornfield (1978) commented on these issues regarding the use of a cluster randomized design by stating that *randomization by cluster accompanied by an analysis appropriate to randomization by individual is an exercise in self-deception and should be discouraged*. Despite the disadvantages and drawbacks of the cluster randomized designs, this design has become increasingly popular in health-related research in the past 20 years. The main reasons are due to other considerations such as ethical issues, feasibility, cost control, and experimental contamination (Donner and Klar, 2000). However, the most important reason is probably experimental contamination. It is difficult to randomly assign half of a cluster such as family, class, or worksite to an intervention and the other half to the concurrent control without introducing possible bias due to contamination. Subjects assigned to different treatment groups within the same cluster are likely to communicate with each other by sharing treatment experience because they are in the same intact social unit. Hence, some of the subjects in the control group may behave like those of the intervention group. As a result, bias is inevitably introduced in the estimation of treatment differences. For example, to evaluate the effectiveness and safety of a screening program using the clinics of general practice as cluster, if investigators or their staff are not blinded to the treatment, they may unconsciously or purposely introduce bias in the evaluation of the subjects in the control group. If the subjects are also not blinded, then it is even at a greater risk of introducing bias because the subjects in the control group may seek screening themselves.

Example 5.3.1 Hutchinson Smoking Prevention Project (HSPP)

The number one cause of preventable death in the United States is cigarette smoking. In addition, more than 90% of adult smokers started smoking by or before age of 20 years old. To encounter this serious problem, in 1983, the U.S. National Cancer Institute initiated a Request for Applications (RFA) on school-based intervention studies for evaluation of the long-term effectiveness of interventions on the prevention of habitual cigarette smoking among youth. In response to this RFA, the Hutchinson Smoking Prevention Project (HSPP; Peterson et al., 2000) was proposed by the Fred Hutchinson Cancer Research Center (FHCRC) to answer the following scientific question:

To what extent can the grade 3–12 HSPP school-based tobacco use prevention intervention deter tobacco use, by both girls and boys, throughout and beyond high school?

The school-based intervention program has become increasingly popular because a school setting is a promising venue for reaching youth with health promotion interventions. However, evaluation of the intervention on smoking prevention program must follow the

following principles of the design and analysis of randomized clinical trials (Peterson et al., 2000):

- 1.** The use of school or school district as the experimental unit for randomization.
- 2.** Sufficient sample size to achieve an acceptable statistical power under a cluster randomized design in the presence of ICC of outcomes within cluster.
- 3.** High rates of recruitment of schools.
- 4.** Random assignment of interventions.
- 5.** Compliance of the intervention across schools.
- 6.** The avoidance of intervention contamination in the control group.
- 7.** The participation of school districts for the duration of the trial.
- 8.** High rates of follow-up and participation for outcome ascertainment.

Selection of school or school district as the experimental unit was the first issue regarding design that the HSPP had to decide at the planning stage of the trial. However, the use of school as a randomization unit may present various design and methodology deficiencies. The HSPP is a randomized, controlled intervention trial with a 10-year intervention from grade 3 to grade 12 with endpoint data collection at the 12th grade and 2 years post-high school. As a result, the most significant issue is the experimental contamination. If school is selected as the experimental unit, contamination of subjects assigned to the control group can occur through unintended acquisition and implementation of the intervention by control teachers, or by social mixing of subjects from both the intervention and control groups, or via student movement from the control to the intervention group. The student movement from one school to another within the same school district poses a serious problem for contamination. Different schools within the same school district have different treatment assignments. As students make the transition from junior high school to high school, students from the control junior high schools will mix up with those from the junior high schools receiving the intervention. The effect of the intervention based on smoking prevalence will be underestimated. Another problem using the school as the experimental unit is the follow-up of the students and collection of the longitudinal data. As reported by Peterson et al. (2000), 49% of the cohort formed at the 3rd grade was no longer enrolled with their original classmates 10 years later at the 12th grade. As a result, using the participating schools for tracking the students and data collection is not totally appropriate. In addition, cooperation from schools always remains as a challenge for research collaboration.

One of the unique design features for the HSPP trial started in 1984 was the selection of school district as experimental unit. The reasons for the decision are summarized below:

- 1.** Permission of the investigation of a multigrade, sequential intervention that spans the elementary, junior-high, and high-school grades.
- 2.** Minimizing the risk of contamination caused by the contamination by the teachers in an intervention school and those in a control school within the same district.
- 3.** Elimination of within-school mixing of intervention and control students during the follow-up.
- 4.** Randomization by school district for the district-wide method by which school districts usually adopt and implement curricula.

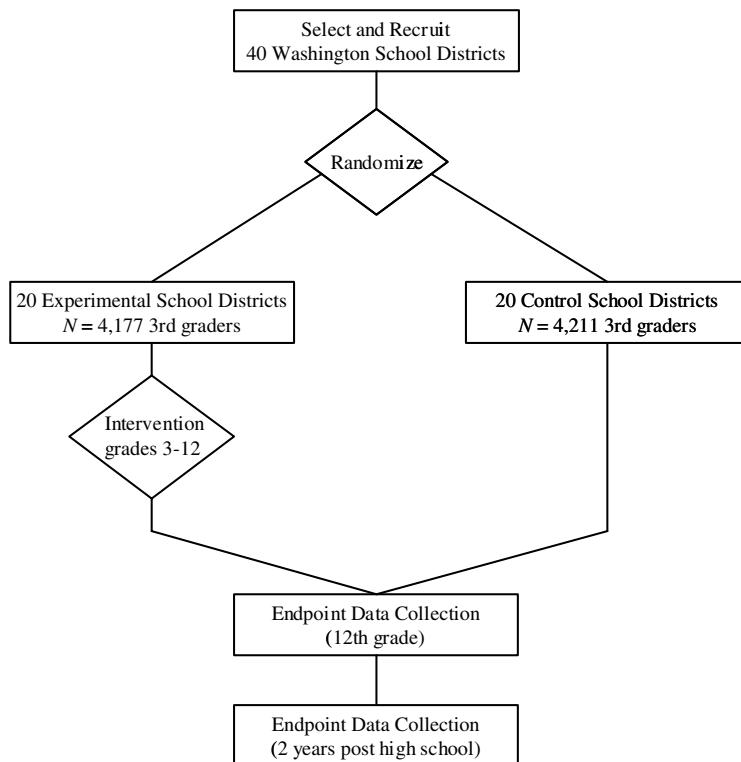


Figure 5.3.1 HSPP experimental design. (Source: Peterson et al., 2000.)

For the purpose of efficient management of the trial, the inclusion criteria for school district are as follows:

1. 50 to 250 students per grade level.
2. Within 200 miles of the FHCRC.
3. With self-contained, stable feeder system consisting of at least elementary school, at least one junior high school and one high school.
4. With grade 3 to 7 attrition of less than 35%.

The matched pairs parallel group design was chosen to take into consideration the important prognostic factors such as tobacco use prevalence, school district size, and location of school district. A total of 40 HSPP school districts met the above inclusion criteria. Twenty were randomized to the intervention program, and the other 20 were randomly assigned to the control program. The diagram of this design is given in Figure 5.3.1.

Example 5.3.2 The WHO Antenatal Care Trial

Donner et al. (1998) and Piaggio et al. (2001) reported an equivalence stratified cluster randomization trial for evaluation of a new antenatal care (ANC) model conducted by

the World Health Organization (WHO) in 53 clinics in Argentina, Cuba, Saudi Arabia, and Thailand. The women in the intervention group received a new antenatal care model consisting of tests, clinical procedures, and follow-up sections, and the women in the control group were given the standard *Western ANC* model currently implemented in these clinics. The primary endpoints for this trial are the proportion with infants of low birth-weight ($<2,500$ g) and the proportion of women with severe maternal morbidity among those with singleton pregnancies. These two endpoints were chosen because they are surrogate variables of perinatal and maternal mortality. The primary hypothesis of the trial is that based on the primary endpoints, the new antenatal care model is equivalent to the standard *Western ANC* model. This trial used the standard average bioequivalence criterion on the original scale (see, e.g., Chow and Liu, 2000) to assess the equivalence between the new and standard ANC models. In other words, the new ANC model is claimed to be equivalent to the standard ANC model if the proportion of lower birth-weight infants (or women with severe maternal morbidity) of the new ANC model is within 20% of that of the western ANC model. Within each country, the clinics were stratified with respect to the size and type of clinics and were randomly assigned to either the new ANC model (27 clinics) or to the standard *Western ANC* model (26 clinics). All pregnant women initiating ANC at these clinics over an average period of 18 months were enrolled to yield a total of 24,678 women recruited. Figure 5.3.2 displayed a diagram of the design for this trial.

5.4 CROSSOVER DESIGNS

A crossover design is a modified randomized block design in which each block receives more than one treatment at different dosing periods. A block can be a patient or a group of patients. Patients in each block receive different sequences of treatments. A crossover design is called a complete crossover design if each sequence contains all treatments under investigation. For a crossover design it is not necessary that the number of treatments in each sequence be greater than or equal to the number of treatments to be compared. We will refer to a crossover design as a $p \times q$ crossover design if there are p sequences of treatments administered at q different dosing periods. Basically a crossover design has the following advantages: (1) It allows a within-patient comparison between treatments, since each patient serves as his or her own control. (2) It removes the interpatient variability from the comparison between treatments. (3) With a proper randomization of patients to the treatment sequences, it provides the best unbiased estimates for the differences between treatments. The use of crossover designs for clinical trials has been much discussed in the literature. See, for example, Brown (1980), Huitson et al. (1982), Jones and Kenward (1989), and Chow and Liu (2000).

For a crossover design the notions of the washout or carryover effects (or residual effects) are important for the analysis of collected clinical data. The washout period is defined as the rest period between two treatment periods for which the effect of one treatment administered at one dosing period does not carry over to the next. In a crossover design the washout period must be long enough for the treatment effect to wear off so that there is no carryover effect from one treatment period to the next. The washout period depends on the nature of the drug. A suitable washout period must be sufficiently long to return any relevant changes that influence the clinical response to the baseline. If a drug has a long half-life or if the washout period between treatment periods is too short, the

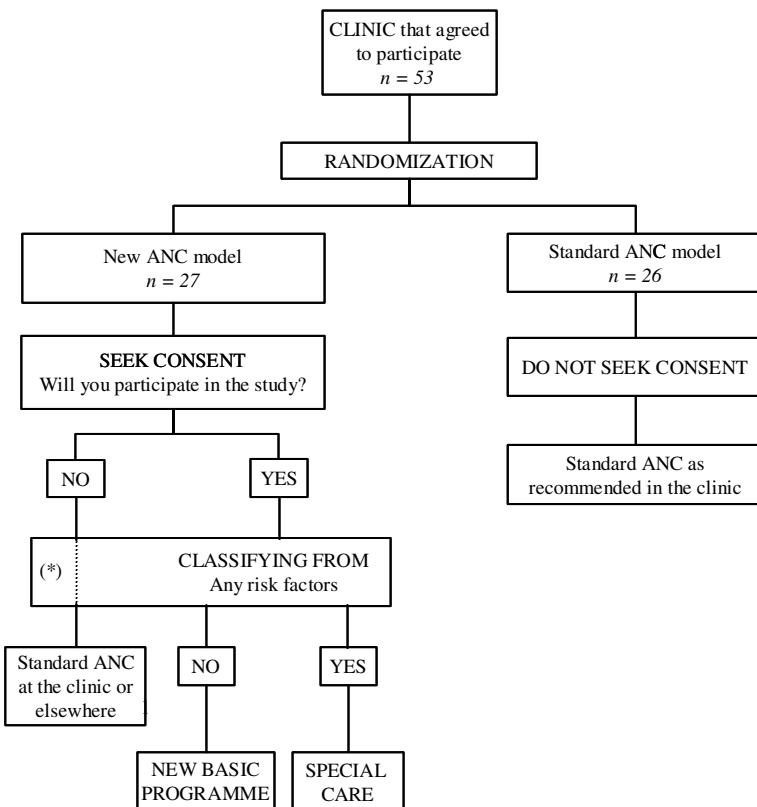


Figure 5.3.2 Study design and subject's flow chart of the antenatal care randomized, controlled trial. Antenatal care clinics are the unit of randomization (cluster randomization). (*) Women who did not agree to participate in the trial were asked to provide information needed to complete a “classifying” for baseline descriptive purposes only. (Source: Piaggio et al., 2001.)

effect of the drug might persist after the end of dosing period. In this case it is necessary to distinguish the difference between the drug effect and the carryover effects. The direct drug effect is the effect that a drug product has during the period in which the drug is administered, while the carryover effect is the drug effect that persists after the end of the dosing period.

Note that crossover designs may be used in clinical trials in the following situations where (1) objective measures and interpretable data for both efficacy and safety are obtained, (2) chronic (relatively stable) disease are under study, (3) prophylactic drugs with relatively short half-life are being investigated, (4) relatively short treatment periods are considered, (5) baseline and washout periods are feasible, and (6) an adequate number of patients for detection of the carryover effect with sufficient power that accounts for expected dropouts is feasible or extra study information is available to rule on the carryover effect. Dubey (1991) also emphasizes that appropriate analyses which can reflect the study design must be carried out when using crossover design in clinical trials.

Higher-Order Crossover Designs

The most commonly used crossover design for comparing two treatments (denoted by A and B) is a two-sequence two-period crossover design. We will refer to this design as a standard 2×2 crossover design, which is sometimes denoted by (AB, BA). For a standard 2×2 crossover design, each patient is randomly assigned to receive either sequence AB or sequence BA at two dosing periods. In other words, subjects within sequence AB (BA) receive treatment A (B) at the first dosing period and treatment B (A) at the second dosing period. The dosing periods of course are separated by a washout period of sufficient length to wear the effect due to the drug received in the first period. An example of a standard 2×2 crossover design is illustrated in Figure 5.4.1. Note that in the crossover design, the number of the treatments to be compared does not necessarily have to be equal to the number of periods. One example is a 2×3 crossover design for comparing two treatments as illustrated in Figure 5.4.2. In this design there are two treatments but three periods. Patients in each sequence receive one of the treatments twice at two different periods.

When the carryover effects are present, a standard 2×2 crossover design may not be desirable because of potential confounding effects. For example, the sequence effect, which cannot be estimated separately, is confounded (or aliased) with the carryover effects. If the carryover effects are unequal, then there exists no unbiased estimate for the direct drug effect from both periods. In addition the carryover effects cannot be precisely estimated because they can only be evaluated based on the between subject comparison. Furthermore the intrasubject variability cannot be estimated independently and directly from the observed data because each subject receives either treatment A or treatment B only once during the study. In other words, there are no replicates for each treatment within each subject. To overcome the above undesirable properties, a higher-order crossover design is usually considered (Chow and Liu, 1992 and 2000). A higher-order crossover design is defined as a crossover design in which either the number of periods is greater than the number of treatments to be compared or the number of sequences is greater than the number of treatments to be compared. There are a number of higher-order crossover designs available in literature (Kershner and Federer, 1981; Laska, Meinser and Kushner, 1983; Laska and Meinser, 1985; Jones and Kenward, 1989). These designs, however, have their own advantages and disadvantages. An in-depth discussion can be found in Jones and Kenward (1989) and Chow and Liu (1992, 2000).

Table 5.4.1 lists some commonly used higher-order crossover designs. The design (AA, BB, AB, BA) is known as Balaam's design (Balaam, 1968). It is the optimal design in the

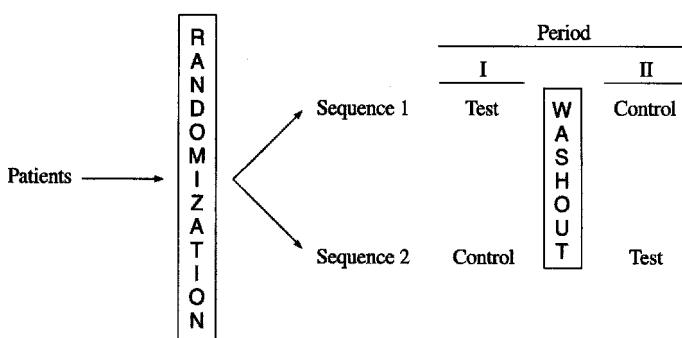
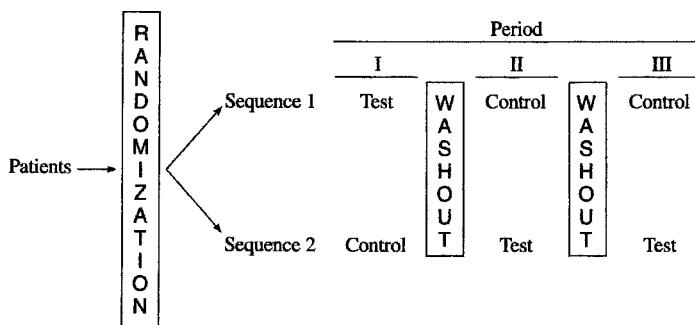


Figure 5.4.1 Standard two-sequence, two-period crossover design.

**Figure 5.4.2** Two-sequence dual crossover design.

class of crossover designs with two periods and two treatments. This design is formed by adding two more sequences (sequences 1 and 2) to the standard 2×2 crossover design (sequences 3 and 4). These two augmented sequences are AA and BB. With additional information provided by the two augmentable sequences, not only can the carryover effects be estimated using the within-subject contrasts but the intrasubject variability for both treatments can also be obtained because there are replicates for each treatment within each subject. Design (ABB, BAA) is the optimal design in the class of crossover designs with two sequences, three periods, and two treatments. It can be obtained by adding an additional period to the standard 2×2 crossover designs. The treatments administered in the third period are the same as those in the second period. This type of designs is also known as the extended-period (extra-period) design. Note that this design is made up of a pair of dual sequences ABB and BAA, and hence it is also known as a two-sequence dual design. Two sequences whose treatments are mirror images of each other are said to be a pair of dual sequences. Jones and Kenward (1989) point out that the only crossover designs worth considering are those made up of dual sequences. Design (AABB, BBAA) is a *doubled* standard 2×2 crossover design (AB, BA). It is usually referred to as a replicated design. Liu (1995b), Liu and Chow (1995), and Chow (1996; 1999) indicate that a replicated design

Table 5.4.1 Higher-Order Crossover Designs

I. Balaam's design
AA
BB
AB
BA
II. Two-sequence dual design
ABB
BAA
III. Doubled (replicated) design
AABB
BBAA
IV. Four-sequence design
AABB
BBAA
ABBA
BAAB

is useful in assessments of bioequivalence between drug products, especially for assessment of individual bioequivalence (Chow et al., 2002) and population bioequivalence (Chow et al., 2003a). Design (AABB, BBAA, ABBA, BAAB) is an optimal design in the class of the crossover designs with four sequences, four periods, and two treatments. It is also made up of two pairs of dual sequences (AABB, BBAA) and (ABBA, BAAB). Note that the first two periods are the same as those in Balaam's design and that the last two periods are the mirror image of the first two periods. This design is much more complicated than designs (AA, BB, AB, BA) and (ABB, BAA), though it produces the maximum variance reduction for both the direct drug effect and the carryover effects among the designs considered.

Williams Designs

When there are more than two treatments to be compared, a complete crossover design becomes much more complicated and may not be of practical interest based on the following considerations:

1. Potential residual effects make the assessment of efficacy and/or safety almost impossible.
2. It takes longer to complete the study.
3. Patients are likely to drop out if they are required to return frequently for tests.

Williams design could be a useful alternative. In this section, for simplicity, we will restrict our attention to those designs in which the number of periods equals the number of treatments to be compared. For comparing three treatments, there are a total of three possible pairwise comparisons between treatments: treatment 1 against treatment 2, treatment 1 against treatment 3, and treatment 2 against treatment 3. It is desirable to estimate these pairwise differences between treatments with the same degree of precision. In other words, it is desirable to have equal variances for each pairwise difference between treatments. Designs with this property are known as variance-balanced designs. It should be noted that in practice, variability associated with the selected design can vary from design to design. Thus an ideal design is the one with the smallest variability such that all pairwise differences between formulations can be estimated with the same and possibly best precision. However, to achieve this goal, the design must be balanced. A design is said to be balanced if it satisfies the following conditions (Jones and Kenward, 1989):

1. Each treatment occurs only once with each subject.
2. Each treatment occurs the same number of times in each period.
3. The number of patients who receive treatment i in some period followed by treatment j in the next period is the same for all $i \neq j$.

Under the constraint of the number of periods (p) being equal to the number of formulations (t), balance can be achieved by using a complete set of *orthogonal Latin squares* (John, 1971; Jones and Kenward, 1989). When the number of treatments to be compared is large, more sequences and consequently more patients are required. This, however, may not be of practical utility. A more practical design has been proposed by Williams (1949). We will refer to this as Williams design. Williams design possesses balance property and requires fewer sequences and periods. The algorithm for constructing a Williams design with t periods and

Table 5.4.2 Williams Designs

I.	Williams's design with three treatments
	ACB
	BAC
	CBA
	BCA
	CAB
	ABC
II.	Williams's design with four treatments
	ADBC
	BACD
	CBDA
	DCAB

t treatments can be found in Jones and Kenward (1989) and Chow and Liu (2000). Table 5.4.2 gives Williams designs for comparing three and four treatments. It can be seen from Table 5.4.2 that Williams design requires fewer sequences in order to achieve the property of variance balance as compared to the complete set of orthogonal Latin squares design. For example, for comparing four treatments, Williams design only requires 4 sequences, whereas a complete set of 4×4 orthogonal Latin squares requires 12 sequences.

Balanced Incomplete Block Design

When there are a large number of treatments to be compared, a complete crossover design may not be feasible. Although a Williams design can be used, it can take a long time to complete. In practice, it is desirable to complete the study in a short period of time. In this case it is desirable that the number of periods is fewer than the number of treatments to be compared. Therefore it is suggested that a randomized incomplete block design be used. An incomplete block design is a randomized block design in which not all treatments are present in every block. A block is called incomplete if the number of treatments in the block is less than the number of treatments to be compared. When an incomplete block design is used, it is recommended that the treatments in each block be randomly assigned in a balanced way so that the design will possess some optimal statistical properties. This kind of design is referred to as a balanced incomplete block design. A balanced incomplete block design is an incomplete block design in which any two treatments appear together an equal number of times. Table 5.4.3 gives two examples of balanced incomplete block designs for comparing four treatments with two periods and three periods, respectively.

Note that a balanced incomplete block design possesses some good statistical properties. For example, unbiased estimates of treatment effects are available and the difference between the effects of any two treatments can be estimated with the same degree of precision.

Examples of Crossover Design in Clinical Trials

To illustrate the use of crossover designs in clinical trials, we again consider the clinical development of Glucophage. Table 5.4.4 lists three studies that have investigated the effects of Glucophage on lipids and other risk factors of cardiovascular disease.

Table 5.4.3 Balanced Incomplete Block Designs

I.	Four treatments with two periods
	AB
	BC
	CD
	DA
	AC
	BD
	DB
	CA
	AD
	DC
	CB
	BA
II.	Four treatments with three periods
	BCD
	CDA
	DAB
	ABC

The objective of the study conducted by Chan et al. (1993) was to compare the metabolic and hemodynamic effects of Glucophage and Glibenclamide in normotensive NIDDM patients. After a two-week run-in period on dietary treatment alone, 12 Chinese normotensive patients with uncomplicated NIDDM were randomized to receive either Glucophage or Glibenclamide for four weeks before being crossovered to the alternative treatment for an additional four weeks. Their metabolic and hemodynamic indices including cardiac output estimation by impedance cardiography were measured at the baseline and at the end of each treatment. The results indicate that at comparable degrees of glycemic control, Glucophage had the following beneficial effects compared with Glibenclamide: (1) greater weight loss (body mass index, -0.58 kg/m^2 vs. -0.12 kg/m^2), (2)

Table 5.4.4 Examples of Crossover Design

Study	Purpose	Sample Size	Duration (Period 1 + Washout + Period 2)	Primary Endpoint
Chan et al. (1993)	Effects on lipids	12	4 wk + 0 wk + 4 wk	Metabolic and Hemodynamic indices
Nagi and Yudkin (1993)	Effects on lipids and risk factors for cardiovascular disease	27	12 wk + 2 wk + 12 wk	Insulin resistance Glycemic control Cardiovascular risk
Elkeles (1991)	Effects on lipids	35	3 mo + 6 wk + 3 mo	Serum lipids Lipoproteins Blood glucose Glycosylated hemoglobin

greater decrease in total cholesterol (-0.7 mM vs. -0.2 mM), and (3) greater decrease in diastolic blood pressure (-12.9 mmHg vs. -6.8 mmHg). In conclusion Chan, et al. (1993) indicates that the tendency to greater peripheral resistance with Glibenclamide and to lower diastolic blood pressure with Glucophage may bear on the development of hypertension in normotensive patients who are receiving the long-term treatment for NIDDM.

For another application of the crossover design, Nagi and Yudkin (1993) investigated the effects of Glucophage on glycemic control, insulin resistance, and risk factors for cardiovascular disease in NIDDM patients with different risks of cardiovascular disease. The study was conducted as a randomized double-blind placebo-controlled crossover design on 27 patients. Glucophage was administered for a total of 12 weeks, and the dose was increased stepwise from 850 mg once daily for one week to 850 mg twice daily for five weeks and to 850 mg three times daily for another six weeks. The baseline assessment took place on the day of inclusion in the study, and a similar assessment took place after 12 weeks of therapy (phase 1). After a washout period of two weeks, patients were reassessed as at entry into the trial and crossed over to the alternative treatment (phase 2). The patients were reassessed finally at the end of phase 2. The results indicated Glucophage reduced fasting plasma glucose levels by 3.08 mM , enhanced insulin sensitivity by 4.0%, and diminished triglyceride levels by 0.2 mM , total cholesterol levels by 0.52 mM , and low-density lipoprotein cholesterol levels by 0.4 mM . Nagi and Yudlin (1993) conclude that Glucophage therapy improves glycemic control by diminishing insulin resistance, enhances lipid and lipoprotein profiles, ameliorates other risk factors for cardiovascular disease independently of weight loss or improved glycemic control, and may therefore have utility in long-term reduction of coronary artery disease risk among patients with NIDDM.

Elkeles (1991) also conducted a three-month crossover trial on 35 patients with poorly controlled NIDDM to investigate the effects of Glucophage and Glibenclamide on body weight, blood glucose control, and serum lipoproteins. After six weeks of a diet that did not achieve adequate diabetic control, patients were randomized to receive either Glibenclamide 5 mg daily or Glucophage 500 mg twice a day. The dose was increased to achieve a fasting blood glucose level of 6 mmol/L or less, up to a maximum of 15 mg Glibenclamide or 3 g Glucophage daily. After three months the treatment was stopped and after six weeks of diet only again, patients were crossed over to receive the other drug. Before and after three months of treatment, blood samples were taken for serum lipids, lipoproteins, blood glucose, and glycosylated haemoglobin. Elkeles reports that Glucophage diminished HbA_{1C} levels by 2.05% and the Glibenclamide by 1.51%. Glucophage also significantly reduced both total cholesterol and low-density lipoprotein cholesterol. Elkeles points out that the improvement in HbA_{1C} levels as well as in total cholesterol and low-density lipoprotein cholesterol reverted when Glucophage was withdrawn for six weeks. Therefore it was concluded that by reducing total cholesterol and low-density lipoprotein cholesterol over the long term, Glucophage can improve the coronary artery disease risk profile independently of its effect on glucose homeostasis.

Example 5.4.1 Joint Use of Parallel Group Design and Crossover Design

As little or no information regarding the interaction between diet and statins (3-hydroxy-3-methylglutaryl coenzyme A [HMG-CoA] reductase inhibitors) is available in the literature, Jula et al. (2002) reported a randomized, controlled crossover trial to evaluate the separate and combined effects of diet and simvastatin therapy on serum levels of lipids, lipoproteins,

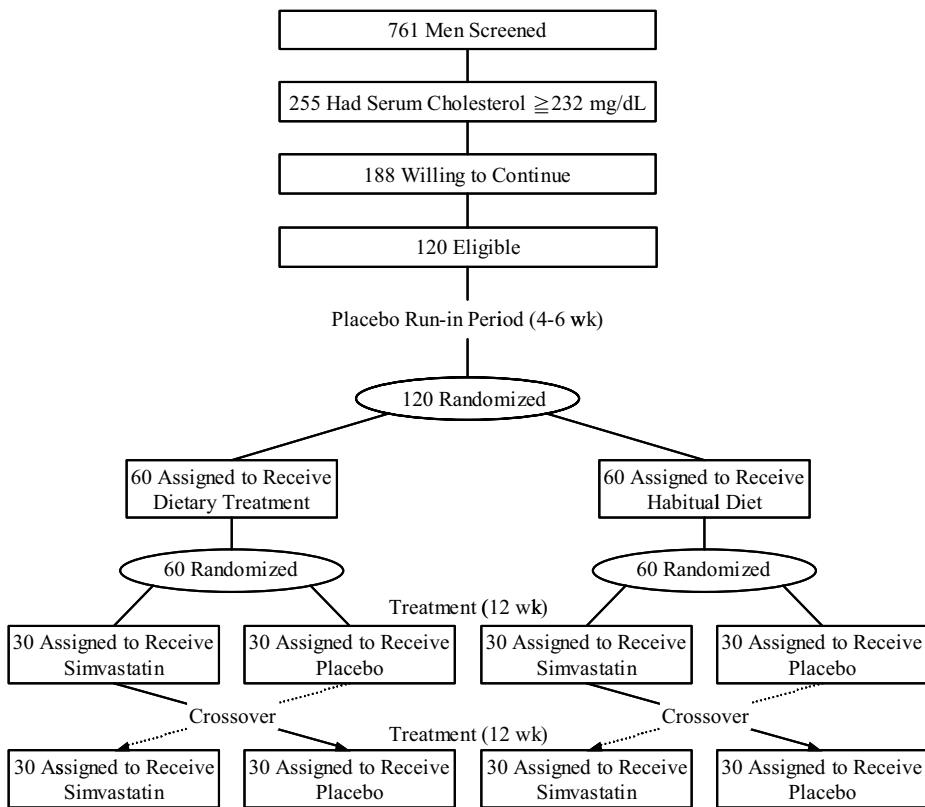


Figure 5.4.3 Joint application of parallel-group and crossover designs. (Source: Jula et al., 2002.)

antioxidants, and insulin. After a 4- to 6-week placebo run-in period, 120 previously untreated hypercholesterolemic men with a fasting serum cholesterol greater than 232 mg/dL were first randomized in a parallel group design with a 1:1 ratio to a habitual diet or dietary treatment group. Within each of these two groups, a standard two-sequence, two-period (2×2) crossover design was employed to randomized subjects to receive simvastatin (20 mg/d) or its matching placebo for 12 weeks in a double-blind fashion. The dietary treatment is a weighted, modified, Mediterranean-type diet. The main goals of the dietary treatment were to reduce energy intake from saturated plus trans-unsaturated fats to no more than 10% by replacing them partly with monounsaturated and polyunsaturated fats rich in omega-3 fatty acids and to increase intake of fruits, vegetables, and dietary fiber. As a pilot study with 20 subjects did not show any evidence of period or carryover effect, no washout period was included between two treatment periods. The diagram of the design for this study is presented in Figure 5.4.3. This study employed a two-group parallel design and the standard 2×2 crossover design to investigate the joint effects of statins and diet. The two-group parallel design in this study is to compare two diets whereas the effects of statins versus placebo were investigated through the standard 2×2 crossover design. This is also an example of factorial trials, which will be discussed later in Chapter 7.

5.5 TITRATION DESIGNS

For phase I safety and tolerance studies, Rodda et al. (1988) classify traditional designs as follows:

1. Rising single-dose design.
2. Rising single-dose crossover design.
3. Alternative-panel rising single-dose design.
4. Alternative-panel rising single-dose crossover design.
5. Parallel-panel rising multiple-dose design.
6. Alternative-panel rising multiple-dose design.

As indicated in Chapter 1, phase I studies are usually conducted in young, healthy male volunteers. The purpose of phase I studies is to obtain initial appraisement of drug safety through the evaluation of vital signs, physical health, and adverse events and frequent assessments of hematology, blood chemistry, and urine samples. The above designs are commonly employed in phase I safety and tolerance studies to efficiently provide the data that can be analyzed for generating hypotheses rather than for making definitive inference.

In medical practice, if the study medicine is intended for cancer or some life-threatening diseases, it may not be ethical to conduct phase I safety and tolerance studies on normal volunteers due to potential toxic or fatal effects. In addition results from animal studies provide little information regarding the therapeutic range for possible efficacy with tolerable safety. Due to the special characteristics of cancer patients and toxic profiles of cancer treatments, designs for cancer clinical trials require special considerations, which will be addressed in Chapter 6.

Standard Titration Design

One of the objectives of phase II clinical trials is to sufficiently characterize the dose-response relationship, which is the most frequently asked question at the FDA Advisory Committee's meeting. The dose-response relationship defines the therapeutic range of a test drug. The therapeutic range is usually referred to as the dosage range between the minimal effective dose (MED) and the maximum tolerable dose (MTD). The MED is the lowest dose above which the efficacy of the test drug is clinically superior to and statistically significant from that of the placebo. The MED is usually demonstrated for at least one primary clinical endpoint. If the range between the MED and MTD is large, then the test drug is said to have a wide *therapeutic window*. If the MED is very close to MTD, then the test drug is considered to have a narrow therapeutic window.

In practice, although preliminary short-term safety information is available after the completion of phase I clinical trials, information on the dose-response relationship and the safety profile with respect to a moderate or long-term exposure of patients to the test drug is usually unknown at the initiation of phase II clinical trials. In order to conform to real clinical practice and to expose the patients only to the amount of dose they need, it is suggested that a titration design be used to provide a conservative and cautious approach in investigating the dose-response relationship, which can in turn be used as a preliminary estimate of the therapeutic window of the drug.

The ICH E4 guideline describes four different designs for assessing dose response: parallel dose-response design, crossover dose-response design, force titration design, and optional titration design (ICH E4, 1994). Parallel and crossover dose-response designs are special applications of parallel group and crossover designs to evaluation of dose-response relationship, respectively. Therefore, although these two designs have been addressed in Sections 5.2 and 5.4, they will be further discussed in this section. Force titration is a variation of the force escalation design that will be described later in this section. On the other hand, the optional titration design is a modified titration design with a concurrent placebo-controlled group where subjects are titrated until they reach a well-characterized favorable or unfavorable response. In this section, both titration and force escalation designs will be introduced along with their variations.

A traditional titration design is also a dose-escalation study with a set of predetermined dose levels and prespecified criteria of responders or nonresponders. A titration design starts with a placebo washout phase during which the previous medications are stopped and the placebo is administered to the patients. At the end of the placebo washout period, the baseline clinical measurements are established. Then, all the patients start with the same lowest dose. A patient is considered a responder and continues to receive the same dose for the duration of the trial if he or she meets the prespecified criteria at the end of the first dosing period. If a patient fails to meet the prespecified criteria at the end of the first dosing period, the dose of the patient is then titrated up to the next higher dose provided that the patient can satisfactorily tolerate the drug. The titration process continues until all dose levels are exhausted. A graphical presentation of a typical titration design is given in Figure 2.4.1. The determination of a responder is usually based on some objective physiological measurements. In practice, the duration of each dosing period is usually chosen to be of sufficient length so that the stabilization of the selected physiological measurements can be achieved to define the titration process.

For the analysis of the data from titration studies, several methods have been proposed in the literature. For example, Chuang (1987) considers the life-table method utilized along with a logistic linear dose-response model. As an alternative, Shih et al. (1989) and Chuang-Stein and Shih (1991) propose the method of EM algorithm. Temple (1982), however, points out that for any titration study the treatment effect is confounded with time. In addition, the primary clinical measurement used to define the titration process might be highly correlated with other efficacy and/or safety endpoints. As a result the missing mechanism for these clinical endpoints are treatment related and are not at random. Consequently the methods for use of longitudinal data such as generalized estimating equations (GEE) proposed by Liang and Zeger (1986) are not appropriate (e.g., see Chuang-Stein, 1993). Therefore, a limitation of titration designs is their lack of valid statistical methodologies for the analysis of continuous or ordered categorical data in adequately characterizing the dose-response relationship of a test drug.

Note that there are other variations associated with a titration design that make statistical analysis even more complicated. An example for such variation is described below. An open-label phase II clinical trial with three titration groups was conducted before availability of the results from CAST to obtain preliminary information of dose-response relationship for a new class IC anti-arrhythmic agent in treatment of patients with ventricular ectopy. Each titration group had a dosing regimen of the four ascending doses as displayed in Table 5.5.1. The first group consists of dose levels of a placebo and 50 mg, 75 mg, and 125 mg t.i.d. (every eight hours) of the drug, the second group includes a placebo, 50 mg, 100 mg, and 150 mg, t.i.d. of the drug, while the third group contains a placebo, 100 mg,

Table 5.5.1 Dose Levels and Number of Patients at Each Dosing Period of the Titration Design for a Class IC Antiarrhythmic Agent

Group	Dosing Period			
	0	1	2	3
A	P(11)	50 (11)	75 (9)	125 (9)
B	P(15)	50 (15)	100 (14)	150 (12)
C	P(16)	100 (16)	150 (14)	200 (9)

Note: P = placebo; the numbers outside the parentheses are dose levels in mg, and the numbers within the parentheses are the number of patients entered each dosing interval.

150 mg, and 200 mg t.i.d. of the drug. The duration of each dosing period is three days, and the patients were assigned to the three groups sequentially. The objective clinical endpoints are derived from the 24-hour Holter recording processed at a central facility. They were:

1. Hourly average of total ventricular premature contraction (VPC) counts.
2. Hourly average of single VPC counts.
3. Sum of couplets over a 24-hour period.
4. Sum of VT runs over a 24-hour period for which the pulse rate was greater than or equal to 100 beats per minutes.

A responder for an adequate suppression of ventricular ectopy is defined as (1) at least a 80% reduction in average total VPC count per hour compared to the baseline value obtained at the end of the placebo dosing period and at least a 90% reduction in the number of events of repetitive forms (couplets or nonsustained ventricular tachycardia, NSVT) or (2) at least a 90% reduction in the number of NSVT provided the number of NSVT is 10 or more during the placebo period. Patients may be withdrawn from the study according to the following efficacy and safety criteria. Efficacy criteria for worsening of ventricular ectopy is defined as either (1) $\ln(Y) \geq 3.118 + 0.646 \ln(X)$, where Y is the hourly average total VPC count at the end of each dosing period for the active treatment and X is the baseline average total VPC hourly count at the end of placebo period, or (2) an increase to 50 or more runs of VPCs if the number of runs over the 24-hour period during the placebo period is less than five or a tenfold increase in the number of runs and if it is at least five during the placebo period. The safety criteria include (1) the QRS interval being at least 180 ms, or (2) an increase of at least 40% in the corrected QC interval compared to the baseline value of the placebo period or a QC greater than 550 ms.

The dose for patients in each group was titrated upward every third day within each dosing group until an adequate suppression of ventricular ectopy according to the above criteria or patients withdrew from the study if they met either criteria for worsening of ventricular ectopy or safety criteria. It should be noted that each dosing group after the placebo period actually is a parallel group design. For example, the last dosing period is in fact compared to three parallel dosing groups: 125 mg, 150 mg, and 200 mg. On the other hand, within each group, comparison among doses is made within each patient. Consequently as an example, a comparison between 50 and 100 mg consists in a comparison between group A and C during the first dosing period and the comparison between 50 and 100 mg within the patients in group B. In addition to the complexity of the design of this study, it is also observed that (1) the study is an open-label study with a nonrandom group assignment,

(2) the three dose titration groups have overlapping doses, and (3) possible confounding dose effects and carryover effects exist because of no washout period between the dosing periods. Therefore it is suggested that the conclusion and/or interpretation of the results based on inferential statistics for comparisons among doses be drawn with extremely caution.

Another issue regarding the interpretation of the information of the dose-response relationship from a titration design is the overestimation of the necessary dose. For example, the results of early titration studies may suggest that a dose of 600 mg per day or more is necessary for the cardio-selective beta-blocker atenol in the effective reduction of blood pressure. However, subsequent parallel-group, placebo-controlled studies demonstrate that a dose above 100 mg per day has no additional effect in the reduction of blood pressure. To overcome this problem, a titration design with a parallel placebo current control may be useful. For example, a phase II clinical trial was conducted to obtain initial dose-response information of an angiotensin-converting enzyme (ACE) inhibitor captopril. Patients are randomized to either the active treatment group or placebo concurrent group. Then the titration process is performed within each group in five dosing periods with five predetermined doses 0, 25 mg, 50 mg, 100 mg, and 150 mg t.i.d. (Temple, 1982). This design is illustrated in Table 5.5.2 with the mean diastolic blood pressure (mmHg). Within each dosing period, patients in the captopril group received the active drug at the titrated dose level and the patients in the parallel placebo concurrent control received its matching placebo. The criterion for a clinical response was defined as a reduction of diastolic blood pressure below 90 mmHg. This design is in fact a parallel group design with two groups. The study can be conducted in a triple-blind fashion in the sense that not the patients nor the investigator nor the sponsor know the actual treatment that is assigned to the patients, although the titration process with the corresponding doses can be made available to everyone. About 70% of patients were titrated up to 100 and 150 mg. If we only examine the results from the active treatment group as if this study had been conducted as the traditional titration design without a parallel placebo concurrent control, there is a very nice dose-response relationship in reduction of blood pressure from the baseline. In this case the results of the active group gives a wrong impression that the test drug produces a monotone increased response up to 150 mg t.i.d. However, at the same time the parallel placebo concurrent group also presents a sizable placebo effect in the reduction of blood pressure. It turns out that the treatment effect, which is the difference in reduction from

Table 5.5.2 Titration Design with a Parallel Placebo Concurrent Control with Diastolic Blood Pressure (mmHg)

	Dose Level				
	0 mg	25 mg	50 mg	100 mg	150 mg
Captopril					
Observed DBD	110	100	99	96	94
Change from 0		-10	-11	-14	-16
Placebo					
Observed DBD	110	104	104	103	101
Change from 0		-6	-6	-7	-9
Difference in					
Change from 0		-4	-5	-7	-6

Source: Summarized from Temple (1982).

baseline between captopril and placebo, reaches a plateau and remain constant after dosing period 2 during which 50 mg t.i.d. was administered. Consequently, Temple (1982) suggests that 50 mg t.i.d. seems to treat most hypertensive patients well.

Forced Dose-Escalation Design

The design illustrated in Table 5.5.2 is an example of the optional titration design mentioned above. Inclusion of a parallel concurrent placebo control group can correct for spontaneous changes and investigator expectations. As each subject in titration designs receives several different doses, in addition to population average dose-response relationship, individual dose-response information can be obtained. In addition, with a careful planning, the titration design may require fewer subjects than the fixed-dose parallel dose response design and fewer subjects may be exposed to higher doses. However, time and dose are confounded with each other. This problem will become particularly troublesome when one tries to characterize the dose-response relationship for adverse events.

Note that some pharmaceutical agents might induce some undesirable but reversible safety concerns. In addition they may not be efficacious at lower doses. When conducting clinical trials with these agents, we would expect a significant number of dropouts. Therefore it is recommended that a trial with these agents begin very cautiously with a very low dose. In such a trial the criteria for titration process is based on safety rather than efficacy because the drug is unlikely to be effective at lower doses. As a result all patients who do not have the predefined safety problem will be forced to receive the next higher dose in the subsequent dosing period. This type of titration design is called the *forced dose-escalation design*. A typical example for obtaining FDA approval using the forced dose-escalation trials is the approval of Tacrine which is intended for treatment of mild to moderate dementia of the Alzheimer's type. Since Tacrine is known to induce elevation of serum alanine aminotransferase (ALT) above the upper limit of the normal range in 43% to 54% of the patients and around 28% of the patients treated with Tacrine showed an elevation of ALT exceeding three times the upper limit of the normal range, the forced dose-escalation design at six-week intervals was chosen for two adequate well-controlled studies for the approval of the drug. The design of the first adequate well-controlled randomized study is a 12-week trial that consists of two six-week double-blind phases with the placebo current control groups as shown in Table 5.5.3 (Farlow et al., 1992). Patients were first randomized to one of the six sequences. For the double-blind phase I the patients in sequences 1 and 2, 3 and 4, and 5 and 6 received placebo, 20 mg, and 40 mg per day, respectively. However, for the double-blind phase II

Table 5.5.3 Forced Dose-Escalation Design for 12-Week Trial of Tacrine in Alzheimer's Disease

Randomized Sequences	Double-Blind Phase I Week 1 to 6	Double-Blind Phase II Week 7 to 12
1	Placebo	Placebo
2	Placebo	Tacrine 20 mg/d
3	Tacrine 20 mg/d	Tacrine 20 mg/d
4	Tacrine 20 mg/d	Tacrine 40 mg/d
5	Tacrine 40 mg/d	Tacrine 40 mg/d
6	Tacrine 40 mg/d	Tacrine 80 mg/d

Table 5.5.4 Forced Dose-Escalation Design for 30-Week Trial of Tacrine in Alzheimer's Disease

Randomized Groups	Week			
	0–6	7–12	13–18	19–30
1	Placebo	Placebo	Placebo	Placebo
2	40 mg/d	80 mg/d	80 mg/d	80 mg/d
3	40 mg/d	80 mg/d	120 mg/d	120 mg/d
4	40 mg/d	80 mg/d	120 mg/d	160 mg/d

patients in sequences 1, 3, and 5 received the same doses as those in the double-blind phase I, while the doses of the patients in sequences 2, 4, and 6 who could tolerate the doses in double-blind phase I were titrated up to 20 mg, 40 mg, and 80 mg during double-blind phase II, respectively. The second adequate well-control study is a long-term 30-week randomized double-blind trial that consists of three parallel groups for the active drug and a parallel placebo current group with a forced dose-escalation design as shown in Table 5.5.4 (Knapp, 1994). To account for anticipated increase incidence of cholinergic adverse events and dropouts at the highest dose, an unequal randomization with a ratio of 3 : 1 : 3 : 4 for groups 1, 2, 3, and 4, respectively, was used for the assignment of patients to the treatment. Patients who were randomized to group 1 received a placebo throughout the entire study. Patients who randomized to group 2 received 40 mg per day for the first six weeks. If they could tolerate the dose, the doses of those patients were escalated to 80 mg per day for the rest of the study. Patients who were randomized to group 3 received 40 mg per day during the first 6 weeks and then were titrated up to 80 mg per day between week 7 and week 12, and again titrated to 120 mg per day for the rest of the study provided that they could tolerate the dose levels. The escalation process for the patients who were randomized to group 4 was 40 mg, 80 mg, 120 mg, and 160 mg per day for the first, second, and third 6-week dosing periods, and for the rest of 12 weeks, respectively. Both trials consist of parallel groups as well as a dose titration process within each group. Therefore they can provide cross-sectional data for comparison among parallel groups as well as longitudinal data for comparison among doses based on the individual patient. However, for the 12-week trial the analysis was performed separately with the cross-sectional data collected at week 6 for the double-blind phase I and week 12 for the double-blind phase II. The cross-sectional data at week 30 were analyzed for the 30-week study. Unfortunately, due to the lack of adequate statistical tools for inference of treatment effects in the presence of treatment-related withdrawal, longitudinal data were not utilized to provide useful information regarding the titration process, indeed such data can be vital for the application of Tacrine in the treatment of patients with probable Alzheimer's disease.

One of the key characteristics for the traditional titration design is that the titration is a dose-escalation process. In other words, the subjects will receive the next higher dose if they meet some predefined efficacy or safety criteria. However, the dose de-escalation is sometimes employed in the titration design. Goldstein et al. (1998) reported a study for evaluation of oral sildenafil in the treatment of males with erectile dysfunction. A total of 329 males 18 years of age or older with a clinical diagnosis of erectile dysfunction of six months or longer were randomly assigned to receive placebo or 50 mg of sildenafil approximately one hour before sexual activity for 12 weeks. At each follow-up visit, the dose can be

doubled up to 100 mg if there is no adequate clinical response and no safety concern. On the other hand, the dose can also be reduced to 25 mg if there is satisfactory clinical response but with a concern about some adverse events. As the dose of this design can be titrated upward as well as downward, it is referred to as the *flexible dose escalation design*. To maintain the study in a double-blind fashion, each dose consists of three tablets from the same row of blister pack in the following configurations: (1) placebo-placebo-placebo, (2) placebo-placebo-25 mg, (3) placebo-placebo-50 mg, or (4) placebo-50 mg-50 mg. Assessment of the primary efficacy is based on the 15-question International Index of Erectile Function (IIEF, Rosen et al., 1997). At the end of the 12-week titration process, the proportions of the subjects receiving 25 mg, 50 mg, and 100 mg in the sildenafil were 2%, 23%, and 74%, respectively. For the placebo group, the corresponding proportions were 0%, 5%, and 95%. Two-hundred-twenty-five men who completed the 12-week study without any serious adverse events were enrolled to receive open-label sildenafil for an additional 32 weeks.

5.6 ENRICHMENT DESIGNS

Some therapeutic agents are likely to be effective in a specific population of patients who may have an underlying disorder that is responsive to the manipulation of dose levels of the same agent or several different agents. In practice, instead of an unselected group of patients, it is of interest to identify the patients in whom the test agent is likely to be beneficial in the early phase of the trial. This phase of manipulation of dose levels of the same therapeutic agent or test of different agents for identification of patients with drug efficacy is called the *enrichment* phase. The patients with drug efficacy identified at the enrichment phase are then randomized to receive either the efficacious dose of the test agent or the matching placebo. A design of this kind is known as an enrichment design.

An enrichment design usually consists of at least two phases. The first phase is the enrichment phase in which an open-label study with a titration design is conducted to use some primary pharmacologic effects to identify patients with a clinical response. The second phase is usually randomized and double-blind, possibly with a concurrent placebo control to formally and rigorously investigate the effectiveness and safety of the test agents in these patients. The concept of enrichment design is illustrated in three clinical trials in the areas of Alzheimer's disease and arrhythmia.

The first example is a clinical trial conducted in the early stage of development of Tacrine with doses of 40 and 80 mg four times a day in treatment of the patients with probable Alzheimer's disease. As indicated by Davis et al. (1992), the reason for the enrichment design to be selected for this trial is that the clinical, biochemical, and pathological heterogeneity of the disease and clinical experience suggest that not all patients will respond to any single treatment and that those who respond might do so only within a limited dose range. This trial consisted of four phases: a six-week double-blind dose-titration enrichment phase, a two-week placebo baseline phase, a six-week randomized double-blind placebo-controlled phase, and a six-week sustained active phase, as is displayed in Figure 5.6.1. Patients who met the inclusion and exclusion criteria were enrolled into the enrichment phase of the trial which consisted of three titration sequences. Each titration sequence consisted of 3 two-week dosing periods. The dose in each titration sequence was always titrated up from 40 to 80 mg four times a day with a placebo in dosing periods 1, 2, and 3 for the titration sequences 1, 2, and 3, respectively. The patients were randomized into one of the three titration sequences

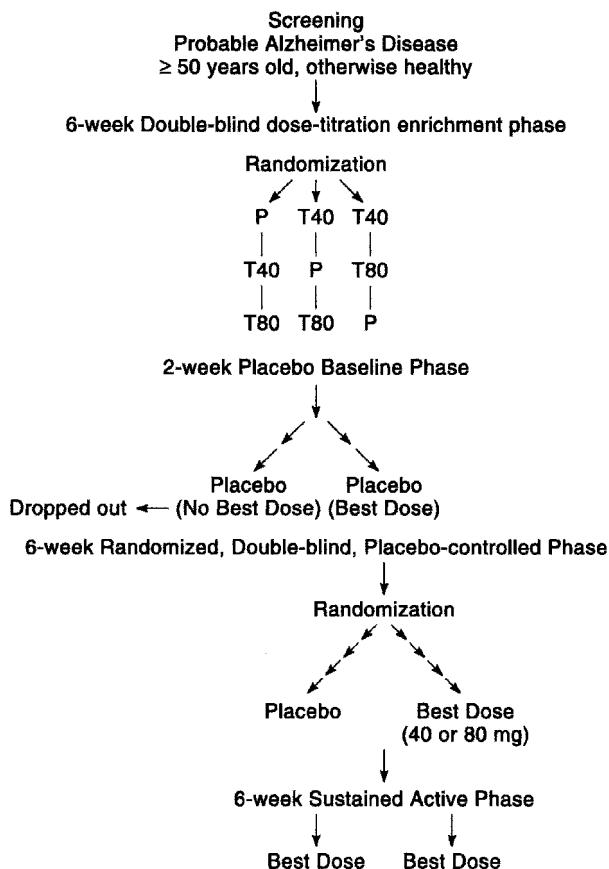


Figure 5.6.1 An enrichment design for tacrine in patients with probable Alzheimer's disease. (Source: Davis et al., 1992.)

which were conducted in double-blind fashion. The potentially therapeutic response for each patient at each dose were then assessed at the end of each two-week dosing period. The *best dose* response for a patient was defined in advance in the protocol as a reduction of at least four points from the screening value in a total score on the Alzheimer's disease assessment scale (ADAS, Folstein et al., 1975) and without intolerable side effects. Then patients with the identified *best dose* were then entered into a two-week placebo baseline period with the hope that this period would be sufficiently long for Tacrine to wear off from the body and for patients to return to the screening pretreatment state with comparable efficacy outcomes. At the end of the two-week placebo baseline phase, the patients with a reduction of at least four point in ADAS during the enrichment phase entered the subsequent six-week randomized double-blind parallel-group, placebo-controlled phase and were randomized with equal probability either to the active Tacrine at their best dose or to the matching placebo. The clinical endpoints measured at the end of the two-week placebo baseline phase served as the baseline for the six-week double-blind phase. Patients who completed the six-week double-blind phase then entered into the sustained active treatment phase.

This study adopted an enrichment design with three titration sequences to identify patients who are likely to respond to Tacrine at a certain dose. After a two-week washout

period, the identified patients were randomized in either Tacrine at their best dose or to placebo concurrent control in a double-blind phase. It, however, should be noted that the fundamental assumption for the use of an enrichment phase is that the *best dose* responses are those obtained when patients are on Tacrine. In practice, it is quite possible that some patients were placebo-responders (i.e., a reduction of at least 4 points on the ADAS scale) who may not respond to Tacrine. In their paper Davis et al. (1992) did not indicate whether there were any placebo-responders in the study. In addition, although there were washout periods between the two-week dosing periods in titration sequences during the enrichment phase, the carryover effect could still exist. Thus it is impossible to estimate the treatment difference unbiasedly based on the data from all three dosing periods during the enrichment phase due to the fact that carryover effects are confounded with treatment effects. Based on the first two weeks of treatment, the enrichment design can provide an unbiased comparison between the placebo and the 40 mg. However, the other two-thirds of the information was wasted. It can be seen from the study that the carryover effect was significant, which suggests that a placebo baseline phase of two weeks was not long enough for the patients to return to the pretreatment state at the screening. In addition, since the carryover effects were confounded with the treatment effects, it is likely that the reduction of at least four points on the ADAS for the *best dose* response was in part due to the carryover effect. As a result Davis et al. (1992) admit that, "Failure to restore baseline conditions fully at the end of the washout period after dose titration makes it impossible to calculate the size of drug effect with certainty." Furthermore it is not clear how to distinguish the characteristics of the patients with the *best dose* response from those who failed to produce a *best dose*. This information is extremely important for practicing physicians who are in favor of prescribing Tacrine to patients with probable Alzheimer's disease. Consequently, this study was not used as one of the two adequate well-controlled studies for approval of Tacrine.

The rationale for selecting the enrichment design in one of the trials during the development of Tacrine is that a short-term response to Tacrine is predictive of the long-term efficacy in prevention of the progression of Alzheimer's disease. The same clinical endpoints were used in both the enrichment and double-blind phases for evaluation of Tacrine's effectiveness. In practice, for other therapeutic agents, the real efficacy endpoint is mortality which requires a longer time to observe. Therefore, the short-term efficacy of the agents is assessed by some other objective surrogate endpoints. It is then very important to know whether the short-term efficacy based on the surrogate endpoint is predictive of a hard endpoint such as mortality. As a result, the enrichment design is usually employed for identification of the *short-term* responders at the initial stage followed by the main phase of the long-term study. Examples of this type of trial can be found in the area of arrhythmia such as in the Cardiac Arrhythmia Suppression Trial (CAST) and the Electrophysiologic Study versus Electrocardiographic Monitoring (ESVEM) trial.

CAST is a multicenter randomized placebo-controlled study sponsored by the U.S. National Heart, Lung, and Blood Institute to test the hypothesis whether the suppression of asymptomatic or mildly symptomatic ventricular arrhythmia after myocardial infarction will reduce the rate of death from arrhythmia. The active drugs included three class IC antiarrhythmic agents Encainide, Flecainide, and Morcizine with a placebo concurrent control. Since the objective of the study was to test the predictability of suppression of ventricular arrhythmia based on ventricular premature contractions (VPC) as recorded by the Holter monitor using the active drugs for mortality, an open-label enrichment design with two titration sequences involved with only active drugs was selected for this study. The patients were stratified by left ventricular ejection (<30%) and time between the qualifying

Holter recording and the myocardial infarction (<90 days or \geq 90 days). Patients with an ejection fraction of at least 30% were assigned at random to receive either the sequence Encainide–Morcizine–Flecainide or the sequence Flecainide–Morcizine–Encainide. The reason for including Morcizine is its inferior efficacy in the suppression of VPC as compared to other two active agents. Each drug was tested at two dose levels. The doses of Encainide, Flecainide, and Morcizine were 35 mg and 50 mg t.i.d., 100 mg and 150 mg, b.i.d., 200 mg and 250 mg t.i.d., respectively. Since Flecainide exhibits negative inotropic properties, it was not administered to the patients with an ejection fraction less than 30%. The pre-specified criteria for an adequate suppression of ventricular arrhythmia were (1) a reduction of at least 80% in VPC and (2) a reduction of at least 90% in runs of unsustained tachycardia as measured by 24-hour Holter recording 4 to 10 days after each dose was begun. The titration process for a particular patient was stopped as soon as a drug and a dose were found to yield adequate suppression. The patients whose arrhythmia were adequately suppressed were then randomized to either the *best drug* identified during the enrichment phase or to placebo for a three-year long-term follow-up. A diagram of the study's design is given in Figure 5.6.2.

The results of CAST, which showed an excessive risk of death for patients who received Encainide or Flecainide as compared to placebo, are thoroughly discussed and examined by Ruskin (1989). The enrichment phase of CAST inherited the fundamental flaw of any titration design in confounding the treatment and carryover effects. For the ethical reason of a minimal exposure of patients to the test agents, the titration process must be stopped as soon as a drug and a dose are found to be effective in suppression of VPC. On the other hand, the optimal drug and dose could not be found for a particular patient. The primary endpoint for CAST is the death or cardiac arrest with resuscitation due to arrhythmia.

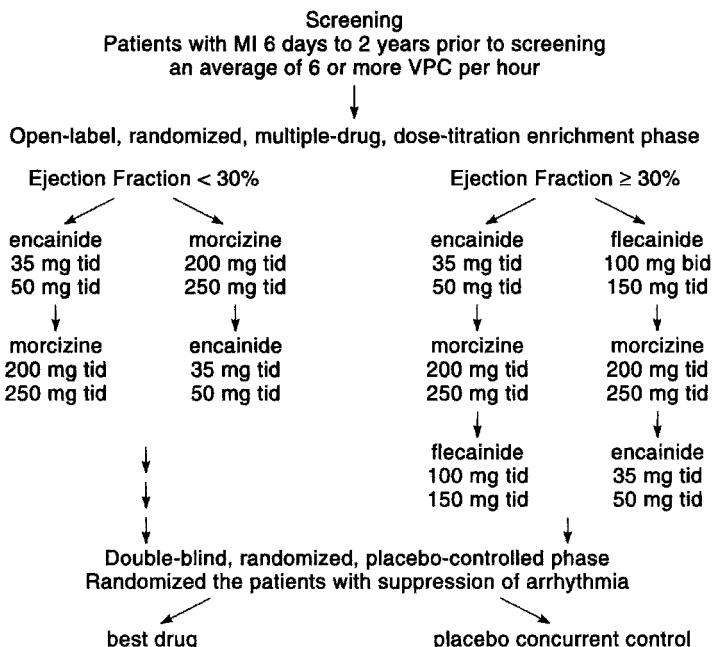


Figure 5.6.2 An enrichment design for cardiac arrhythmia suppression trial (CAST).

However, the analysis for the comparison of mortality rate between antiarrhythmic agents and placebo failed to take into account the occurrence of those events during the enrichment phase because of the lack of an adequate statistical tool for combining the results based on the same primary endpoint from both the enrichment phase and the randomized, placebo-controlled phase. On the other hand, other patients with adequate suppression based on VPC as a therapeutic endpoint might respond very differently from patients with arrhythmia, such as those with ventricular dysfunction or unsustained ventricular tachycardia. This can be seen from the comparison of the mortality rate assumed for sample size determination and the actual mortality rate. The mortality rate of the placebo concurrent control for the sample size determination of CAST over a period of three years was 11%, while the observed mortality of the placebo group was only about 2.2%. As a result the enrichment design in CAST produced a biased estimate of mortality rate for the placebo concurrent control and excluded death in open-label titration process during the enrichment phase in the suppression of arrhythmia for a patient population with a low risk of death.

The results of CAST indicate that the short-term efficacy measured as suppression of VPC based on noninvasive ambulatory electrocardiographic monitoring such as Holter monitor might not be a good predictor where mortality is the endpoint. Others have argued that the failure to induce ventricular tachycardia or fibrillation by some drug assessed by the invasive electrophysiologic study might be a good alternative independent predictor of recurrence of arrhythmia. Consequently, the ESVEM trial sponsored by the U.S. National Institute of Heart, Lung, and Blood was the first large prospective randomized trial conducted to compare the two methods for predictability of long-term recurrence of arrhythmia by the short-term efficacy assessed by the two methods (ESVEM investigators, 1989, 1993; Mason and ESVEM investigators, 1993a, 1993b). For the assessment of predictability, a correlation was obtained using the difference in the recurrence rates of arrhythmia with the short-term efficacy by both methods. An inpatient enrichment phase was elected to identify patients in whom a test drug exhibited a short-term efficacy assessed by either of the two methods. This enrichment design is illustrated in Figure 5.6.3. Patients who met the entry criteria and a 48-hour Holter monitoring and electrophysiologic study criteria were randomized to one of two parallel groups for the two methods in order to assess the short-term drug efficacy. For the assessment of the short-term drug efficacy, the first group employed noninvasive ambulatory electrocardiographic monitoring, while the second group applied the invasive electrophysiologic study. Within each group the patients received up to six arrhythmia agents in a random order until one drug was predicted to be efficacious or until all drugs were tested. A test drug is classified as efficacious when assessed by electrocardiographic monitoring during the inpatient enrichment phase if the following efficacy criteria are met: (1) 70% reduction in mean VPC count, (2) 80% reduction in VPC pair count, (3) 90% reduction in mean ventricular tachycardia counts, and (4) absence of any runs of ventricular tachycardia longer than 15 seconds. Drug efficacy evaluated by the electrophysiologic study during the enrichment phase is defined as failure to induce a run of ventricular tachycardia longer than 15 seconds with V1V2V3 stimulation at the right ventricular apex. If a drug was proved to be efficacious for a patient during the enrichment phase, then he or she was discharged from the hospital for the long-term follow-up with the drug, and the accuracy of the prediction of efficacy was determined during the long-term follow-up. Patients in whom no drugs were proved to be effective during the enrichment phase were not randomized and were withdrawn from the study. However, the vital signs and the recurrence of arrhythmia of the withdrawn patients were monitored.

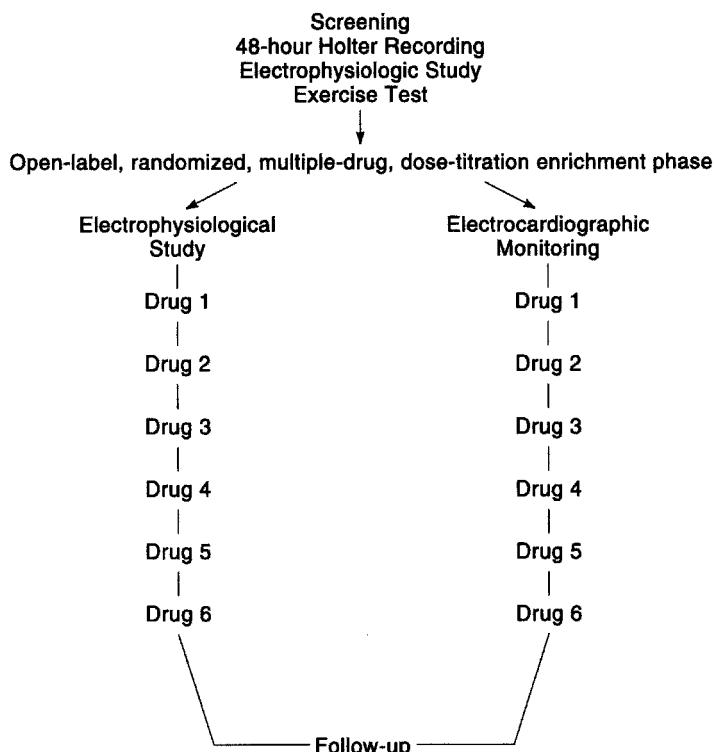


Figure 5.6.3 An enrichment design for electrophysiologic study versus electrocardiographic monitoring (ESVEM). (Source: The ESVEM Investigators, 1989).

The application of an enrichment design in the ESVEM study narrowed the patient population to a very highly selective minority of sustained ventricular tachycardia associated with coronary disease (Ward and Camm, 1993). The patients had to have frequent ventricular ectopic beats by the Holter monitor and inducible ventricular tachycardia or fibrillation by electrophysiological study. Any meaningful clinical inference was severely limited by this constraint. In addition, patients were discharged from the hospital as soon as the first drug proved to be effective. An average of 2.6 (out of 6) drug assessments were performed for each patient. Hence, the patients went to the long-term follow-up on a drug that was effective but not necessarily the optimal one for the patient. Although the investigator did not know what the next drug to be tested was until the proceeding drug failed, he or she was not blinded to the drug currently being tested. Therefore, bias could occur in clinical judgment or in the interpretation of the results. Mason and the ESVEM (1993b) also compared the differences between the active agent sotolol and the other drugs whose efficacy was defined as (1) the drug was tolerable during the enrichment phase, (2) the drug was predicted to be effective, (3) arrhythmia did not occur, and (4) the drug was not discontinued because of an adverse event. However, the results from this analysis (despite its intention-to-treat database) are biased because (1) the assignment of patients to drugs was not at random, (2) patients entered the long-term follow-up at the first drug that met the short-term efficacy criteria, and (3) there was no placebo concurrent control which was proven to be so crucial in CAST.

In summary, an enrichment design is a part of screening process that further restricts the target population to a small selective group. However, sometimes it is still not possible to distinguish this small group from other patients with the same ailment in terms of demographic and other prognostic factors. On the other hand, statistical methods for analysis based on the data from the enrichment phase and the double-blind or primary phase of the trial are not fully developed due to (1) the lack of randomization and (2) different methods of randomization for the enrichment phase. Therefore, the statistical analysis and clinical interpretation for a trial using an enrichment design remain a challenge to both statisticians and clinicians.

5.7 GROUP SEQUENTIAL DESIGNS

One of the unique features and special characteristics of clinical trials is acquisition of experimental units. All experimental units in most *in vitro* experiments, agricultural field trials, and animal trials can be assembled at the beginning of the study, treatments can be applied to all experiment units, and evaluations can be concurrently performed at the same time with respect to a uniform schedule. As mentioned before, however, in clinical trials, the experimental units to which treatment is applied are human subjects. Therefore, unlike other type of experiments, subjects are recruited sequentially over a time interval called the *accrual* period that can range from weeks to months and years. After subjects are screened and their characteristics meet the inclusion and exclusion criteria of the study, they are to receive the assigned (probably randomly) treatment for a prespecified duration and to be evaluated over a follow-up period. The total duration of a clinical trial is hence determined by the accrual and follow-up periods.

In general, clinical trials are longitudinal in nature. Not only are subjects enrolled into a clinical trial in a staggering manner, but also the information generated by the study is accumulated sequentially over time. In addition, the conduct of the trial should be monitored to verify whether it is carried out according to the protocol and it follows the *Good Clinical Practice*. Moreover, for ethical reasons and in the best interest of subjects, a mechanism should be established to terminate the trial before its scheduled completion if it generates convincing evidence of either benefit or harm of the investigational drug. A *group sequential design* provides an assessment of subject outcomes in a group and sequential fashion periodically during the trial rather than on a continuous basis as data from each subject become available. In the past 20 years, various methods of interim analyses (Pocock, 1977; O'Brien and Fleming, 1979; Lan and DeMets, 1983) were developed to assess effectiveness and safety of a drug during the trial adopted a group sequential design. An independent *data and safety monitoring committee* (DSMC) is usually set up to monitor the conduct and safety of the trial and to review or perform the interim analysis arising from a trial with a group sequential design. ICH E9 guideline on *Statistical Considerations for Clinical Trials* stressed that although the group sequential design is in general employed in larger and long-term trials, safety must be monitored in all trials and the need for formal procedures to all early termination for safety reasons should always be considered. Despite the fact that methodologies of group sequential designs and interim analyses have become matured (DeMets and Lan, 1994), new challenges for interim analysis and DSMC are emerging (Dixon and Lagakos, 2000; DeMets, 2000). Different methods of interim analyses will be reviewed in Chapter 10, and the issues associated with group sequential design and DSMC will be addressed in Chapter 12.

Example 5.7.1 Multicenter Automatic Defibrillator Implantation Trial II (MADIT II)

Patients with reduced left ventricular function after myocardial infarction (MI) are at risk of congestive heart failure (CHF) and life-threatening ventricular arrhythmia. Moss et al. (2002) reported the results of the Multicenter Automatic Defibrillator Implantation Trial II that was designed to evaluate the potential survival benefit of a prophylactically implanted defibrillator (in the absence of electrophysiological testing to induce arrhythmia) in patients with a prior myocardial infarction and a left ventricular ejection fraction of 0.30 or less. Over a period of 4 years, a total of 1,232 patients were randomized in a 3:2 ratio to receive an implanted defibrillator (742 patients) or conventional medical therapy (490 patients). The defibrillator was implanted using the standard technique. Every effort was made to achieve defibrillation within a 10-J safety margin. The primary endpoint was all causes death. This trial employed a triangular sequential design (Whitehead, 1997) that was modified for a two-sided alternative and corrected for the lag in obtaining data accrued but not reported before the termination of the trial. This particular sequential design was chosen to allow weekly monitoring with prespecified boundaries to permit early termination of the trial if the defibrillator therapy was found to be superior to, inferior to, or equal to conventional medical therapy. The frequency for interim analysis and monitoring is intensive for this trial as compared to that of usual group sequential design. Hence, it required a careful planning, preparation, administration, and execution of interim analyses and data monitoring. A graphical presentation of paths for this trial is reproduced in Figure 5.7.1. Patients started to enroll into the trial in July 1997. From Figure 5.7.1, an interim analysis on November 13, 2001 showed that the difference in mortality between the defibrillator therapy and conventional medical therapy had reached the prespecified efficacy boundary for superiority. The trial was terminated on November 20, 2001 based on the recommendation of the data and safety monitoring

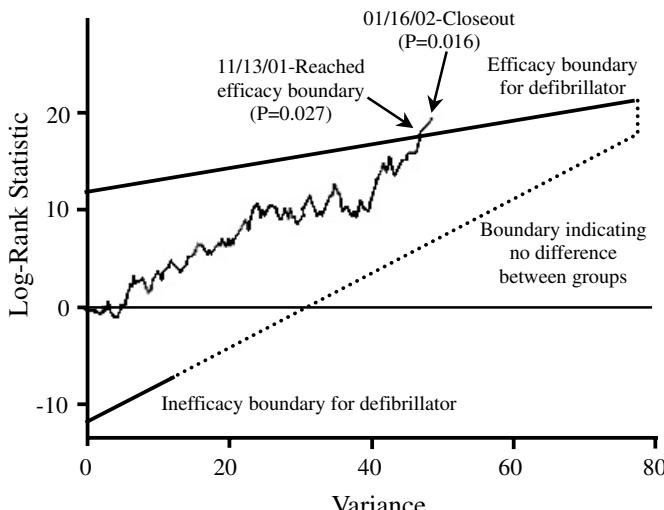


Figure 5.7.1 Sequential monitoring in the triangular design for MADIT II. (Source: Moss et al., 2002.)

committee. Database was released on January 16, 2002 for the final analysis where during the closeout procedure additional survival data and five more deaths that occurred before the stopping dates were uncovered due to a lag in reporting. As a result, the duration of the trial is 4 years and 5 months. It should be noted that a patient was entered into the follow-up period of the study once they received the assigned treatment. The patient's follow-up period, therefore, ranged from 6 days to 53 months with an average of 20 months.

Example 5.7.2 Chemotherapy Regimens for Advanced Non-Small-Cell Lung Cancer

The survival rate of patients with untreated non-small-cell lung cancer at one year is only 10% with a median survival of 4 to 5 months (Rapp et al., 1988). Lung cancer accounts for about one-third of all cancer-related deaths. Over the last 10 years, a number of new agents become available for the treatment of patients with metastatic non-small-cell lung cancer. Although these newer chemotherapy regimens are currently used frequently, fewer comparisons have been made to evaluate the effectiveness among these regimens. Schiller et al. (2002) reported a randomized clinical trial conducted by the Eastern Cooperative Oncology Group (ECOG) to compare the efficacy of three commonly used regimens with that of a reference regimen of cisplatin and paclitaxel. The three regimens included in this trial were cisplatin plus gemcitabine, cisplatin plus docetaxel, and carboplatin plus paclitaxel. The primary endpoint was the overall survival from the date of enrollment. The planned sample size for this study was 300 patients per treatment group over a 30-month period. The ECOG Data Monitoring Committee was responsible for data and safety monitoring as well as interim analyses. This trial employed a group sequential design with a frequency of interim analyses much less intensive than that given in Example 5.7.1. The group sequential design specified two interim analyses and one final analysis of the overall survival when one-third, two-thirds, and all of the anticipated number of deaths had occurred. This trial was executed until the planned number of patients had been enrolled in the study and a total of 1,207 patients were enrolled in the trial from October 1996 to May 1999. The conclusion of this trial is that none of four chemotherapy regimens offered a significant advantage over the others in the treatment of advanced non-small-cell lung cancer.

5.8 PLACEBO-CHALLENGING DESIGN

As indicated earlier, in many clinical trials, a parallel group design alone or a crossover design alone may not be appropriate for evaluation of the safety and efficacy of some drug products. Instead, a combination of a parallel group design or a crossover design with the characteristics of some other designs, such as titration design and enrichment design, may be more appropriate. For example, for evaluation of the efficacy and safety of drug products for treatment of erection dysfunction in male subjects, a design that consists of a titration phase for achieving optimal dose and a crossover active treatment phase with two placebo challenges (i.e., pre- and post-treatment) is often considered. We will refer to the design of this kind as a placebo-challenging design.

For the placebo-challenging design of this kind, subjects are evaluated for eligibility at screening based on medical history, laboratory test, and physical examination. Eligible subjects then enter the dose titration phase. Each subject proceeds a stepwise dose titration until a minimal dose that produces the optimal response is identified. An optimal response is

defined as an erection sufficient to achieve vaginal penetration and lasting from 30 to 80 minutes. At the start of treatment, subjects are required to undergo an in-clinic evaluation of a double-blind placebo-challenge (i.e., subjects will be randomized to receive either the placebo or the active dose at the level identified during the titration). After this first period of in-clinic study, all subjects are to receive a three-month home treatment period at the dose identified during the titration. At the end of three-month treatment, a second in-clinic double-blind placebo-challenging is conducted. Note that at the second placebo-challenging, patients are randomly assigned to receive either the placebo or the active dose. As a result, the above design is in fact the combination of a titration design and a four-sequence, two-period (4×2) crossover design. In other words, during the crossover phase, there are four sequences of treatments, namely, PP, PA, AP, and AA. As an example, for the sequence of AP, subjects are randomized to receive the active dose at the start of the home treatment and are randomly assigned to receive the placebo at the end of the three-month home treatment.

Statistical Model and Inferences

For the placebo-challenging design described above, standard statistical procedures may not be applicable. Chow et al. (2000) studied statistical properties of a placebo-challenging design and developed some new statistical methodology for analysis of data collected from such a design. The statistical methodology is briefly outline below.

Suppose that there are a total of $2n$ subjects in a placebo-challenging design, as described above, qualified subjects are randomly assigned to two treatment groups (i.e., placebo and treatment) in the first and the last periods of the study. Each group consists of n patients, and the assignments in two periods are independent. Thus, $2n$ patients can be classified into the following four groups according to the type of treatments received in two periods:

Group	Treatment	Number of patients
1	PP	n_1
2	AA	n_1
3	PA	n_2
4	AP	n_2

where $n_1 + n_2 = n$. Data will be collected in the first and the last periods. A three-month home treatment period will be given to each patient, and no data will be collected.

Note that the placebo-challenging design is very similar to that of a four-sequence, two-period (4×2) crossover design (Jones and Kenward, 1989). The major difference between the two designs is that in the placebo-challenge study, there is a three-month home treatment between the two periods of placebo/treatment. Consequently, statistical models (parameter specifications) under the two designs are different. For example, in a 4×2 crossover design, one can estimate the period effect, but in a placebo-challenge design, the period effect is confounded with the three-month home treatment effect. As the period effect is usually much smaller than the three-month home treatment effect, we may assume that the former is negligible so that the three-month home treatment effect is estimable.

Let y_{ijk} be the observation from the i th patient in the j period and the k th group. We propose the following statistical model:

$$y_{i11} = \mu + q_1 - \tau + S_{i11} + e_{i11}$$

$$y_{i21} = \mu + q_1 - \tau + \beta - \gamma_1 + \gamma_2 + S_{i21} + e_{i21}$$

$$\begin{aligned}
y_{i12} &= \mu + q_2 + \tau + S_{i12} + e_{i12} \\
y_{i22} &= \mu + q_2 + \tau + \beta + \gamma_1 + \gamma_2 + S_{i22} + e_{i22} \\
y_{i13} &= \mu + q_3 - \tau + S_{i13} + e_{i13} \\
y_{i23} &= \mu + q_3 + \tau + \beta - \gamma_1 - \gamma_2 + S_{i23} + e_{i23} \\
y_{i14} &= \mu + q_4 + \tau + S_{i14} + e_{i14} \\
y_{i24} &= \mu + q_4 - \tau + \beta + \gamma_1 - \gamma_2 + S_{i24} + e_{i24}
\end{aligned}$$

where μ is the overall mean, q_k 's are block (group) effects, $q_1 + q_2 + q_3 + q_4 = 0$, τ is the placebo-treatment effect, β is the three-month home treatment effect (plus the period effect if it is not negligible), γ_1 and γ_2 are interaction effects of two treatment effects (τ and β) and block effects, e_{ijk} 's are independent random errors with mean 0, S_{ijk} 's are random subject effects with mean 0, and the pairs (S_{i1k}, S_{i2k}) are independent. Note that S_{i1k} and S_{i2k} are random effects from the same subject, and therefore, they may be correlated.

Let $\mu_{jk} = E(y_{ijk})$, $j=1, 2$, $k = 1, \dots, 4$, be the group means. Then, a matrix form of the proposed model is

$$\begin{bmatrix} \mu_{11} \\ \mu_{21} \\ \mu_{12} \\ \mu_{22} \\ \mu_{13} \\ \mu_{23} \\ \mu_{14} \\ \mu_{24} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & -1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 & -1 & 1 & -1 & 1 \\ 1 & 0 & 1 & 0 & 1 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & -1 & -1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 & -1 & 0 & 0 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ q_1 \\ q_2 \\ q_3 \\ \tau \\ \beta \\ \gamma_1 \\ \gamma_2 \end{bmatrix} \quad (5.8.1)$$

The effects τ and β can be interpreted as follows. Assume that q_i 's are 0. If there is no interaction $\gamma_1 = \gamma_2 = 0$, then $\mu_{11} = \mu_{13}$ (the mean under placebo in the first period), $\mu_{12} = \mu_{14}$ (the mean under treatment in the first period), $\mu_{21} = \mu_{23}$ (the mean under placebo in the last period), and $\mu_{22} = \mu_{24}$ (the mean under treatment in the last period). When there are interactions, $\mu_{11} = \mu_{13}$ and $\mu_{12} = \mu_{14}$ still hold, but in the last period, $\mu_{21} \neq \mu_{23}$ and $\mu_{22} \neq \mu_{24}$.

Under the model proposed by Chow et al. (2000), some statistical inferences can be obtained as follows.

Let \bar{y}_{jk} be the sample mean based on y_{ijk} 's with fixed j and k . Under model (5.8.1), unbiased estimators of model parameters can be derived by inverting the matrix on the right-hand side of (5.8.1); that is,

$$\begin{bmatrix} \hat{\mu} \\ \hat{q}_1 \\ \hat{q}_2 \\ \hat{q}_3 \\ \hat{\tau} \\ \hat{\beta} \\ \hat{\gamma}_1 \\ \hat{\gamma}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{8} & \frac{1}{8} \\ 1 & -\frac{1}{4} & -\frac{1}{2} & \frac{1}{4} & -\frac{1}{2} & \frac{1}{4} & 0 & -\frac{1}{4} \\ -\frac{1}{2} & \frac{1}{4} & 1 & -\frac{1}{4} & 0 & -\frac{1}{4} & -\frac{1}{2} & \frac{1}{4} \\ 0 & -\frac{1}{4} & -\frac{1}{2} & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & -\frac{1}{4} \\ \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} & -\frac{1}{8} & \frac{1}{8} \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \end{bmatrix} \begin{bmatrix} \bar{y}_{11} \\ \bar{y}_{21} \\ \bar{y}_{12} \\ \bar{y}_{22} \\ \bar{y}_{13} \\ \bar{y}_{23} \\ \bar{y}_{14} \\ \bar{y}_{24} \end{bmatrix} \quad (5.8.2)$$

It follows that (5.8.2) that estimators of τ , β , γ_1 , and γ_2 are based on the individual differences

$$d_{ik} = y_{i1k} - y_{i2k}, \quad j = 1, 2, k = 1, \dots, 4.$$

Although y_{ijk} 's are correlated for the same subject, d_{ik} 's are independent. Let \bar{d}_k be the sample mean based on d_{ik} , $i = 1, \dots, n_l$, where $l = 1$ when $k = 1, 2$, and $l = 2$ when $k = 3, 4$. Then, by (5.8.2),

$$\begin{bmatrix} \hat{\tau} \\ \hat{\beta} \\ \hat{\gamma}_1 \\ \hat{\gamma}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} \\ -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} & -\frac{1}{8} \\ \frac{1}{2} & -\frac{1}{2} & 0 & 0 \\ -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} \bar{d}_1 \\ \bar{d}_2 \\ \bar{d}_3 \\ \bar{d}_4 \end{bmatrix} \quad (5.8.3)$$

As d_{ik} 's are independent, it follows from (5.8.3) that

$$Var(\bar{d}_k) = \frac{\sigma_k^2}{n_l}, \quad k = 1, \dots, 4,$$

where

$$\sigma_k^2 = Var(d_{ik}) = Var(y_{i1k}) + Var(y_{i2k}) - 2Cov(y_{i1k}, y_{i2k}).$$

Consequently,

$$Var(\hat{\tau}) = \frac{1}{16} \left(\frac{\sigma_1^2 + \sigma_2^2}{n_1} + \frac{\sigma_3^2 + \sigma_4^2}{n_2} \right),$$

$$Var(\hat{\beta}) = \frac{1}{64} \left(\frac{\sigma_1^2 + \sigma_2^2}{n_1} + \frac{\sigma_3^2 + \sigma_4^2}{n_2} \right),$$

$$Var(\hat{\gamma}_1) = \frac{\sigma_1^2 + \sigma_2^2}{4n_1}$$

$$Var(\hat{\gamma}_2) = \frac{1}{16} \left(\frac{\sigma_1^2 + \sigma_2^2}{n_1} + \frac{\sigma_3^2 + \sigma_4^2}{n_2} \right).$$

Assume that y_{ijk} 's are normally distributed and $\sigma_k^2 = \sigma^2$ for all k . Then, exact $1 - \alpha$ confidence intervals for τ , β , γ_1 , and γ_2 are given by, respectively,

$$\hat{\tau} \pm \frac{t(\alpha, 2(n-2))\hat{\sigma}}{4} \sqrt{\frac{2}{n_1} + \frac{2}{n_2}}$$

$$\hat{\beta} \pm \frac{t(\alpha, 2(n-2))\hat{\sigma}}{8} \sqrt{\frac{2}{n_1} + \frac{2}{n_2}}$$

$$\hat{\gamma}_1 \pm \frac{t(\alpha, 2(n-2))\hat{\sigma}}{2} \sqrt{\frac{2}{n_2}}$$

$$\hat{\gamma}_2 \pm \frac{t(\alpha, 2(n-2))\hat{\sigma}}{4} \sqrt{\frac{2}{n_1} + \frac{2}{n_2}},$$

where

$$\hat{\sigma}^2 = \frac{(n_1-1)\hat{\sigma}_1^2 + (n_1-1)\hat{\sigma}_2^2 + (n_2-1)\hat{\sigma}_3^2 + (n_2-1)\hat{\sigma}_4^2}{2(n-2)},$$

$\hat{\sigma}_k^2$ is the sample variance based on d_{ik} , $i = 1, \dots, n_i$, and $t(\alpha, 2(n-2))$ is the upper $\alpha/2$ quantile of the t-distribution with $2(n-2)$ degrees of freedom.

If σ_k^2 are different or y_{ijk} 's are not normally distributed, then it is difficult to obtain exact confidence intervals. When n is large, approximate $1 - \alpha$ confidence intervals for τ , β , γ_1 and γ_2 are given by, respectively,

$$\begin{aligned}\hat{\tau} &\pm \frac{Z(\alpha)}{4} \sqrt{\frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{n_1} + \frac{\hat{\sigma}_3^2 + \hat{\sigma}_4^2}{n_2}} \\ \hat{\beta} &\pm \frac{Z(\alpha)}{8} \sqrt{\frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{n_1} + \frac{\hat{\sigma}_3^2 + \hat{\sigma}_4^2}{n_2}} \\ \hat{\gamma}_1 &\pm \frac{Z(\alpha)}{2} \sqrt{\frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{n_1}} \\ \hat{\gamma}_2 &\pm \frac{Z(\alpha)}{4} \sqrt{\frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{n_1} + \frac{\hat{\sigma}_3^2 + \hat{\sigma}_4^2}{n_2}},\end{aligned}\tag{5.8.4}$$

where $Z(\alpha)$ is the upper $\alpha/2$ quantile of the standard normal distribution.

Example 5.8.1 Erectile Dysfunction

A clinical trial was conducted with the placebo-challenging design, as described above, to evaluate the efficacy and safety of a study drug for treating male patients with erectile dysfunction. A total of 120 male patients met the inclusion and exclusion criteria for eligibility. At each placebo-challenging, patients were randomly assigned to receive either the study drug or the placebo. As a result, patients received different sequences of treatments, which are summarized below:

Group	Treatment	Number of patients
1	PP	32
2	AA	32
3	PA	28
4	AP	28

The primary efficacy evaluations in this study was based on the efficacy outcomes of the double-blind placebo challenges using a 3-point erection assessment scale (1 = no response, 2 = partial response, i.e., some enlargement but not sufficient for intercourse, and 3 = full erection). The erection responses are listed in Table 5.8.1.

Table 5.8.1 Data from a Placebo-Challenging Design

Sequence Period	PP		AA		PA		AP	
	1	2	1	2	1	2	1	2
Subject 1	0	0	0	1	0	1	0	1
2	1	0	1	2	0	2	1	2
3	1	1	0	0	1	2	2	2
4	0	2	0	1	1	1	2	1
5	0	1	0	1	1	1	0	1
6	1	2	0	1	0	0	2	2
7	1	2	0	2	0	2	0	1
8	0	0	2	2	2	1	1	1
9	1	1	0	1	1	1	0	1
10	1	1	1	1	1	2	0	1
11	0	0	0	1	0	2	2	2
12	1	2	1	2	1	2	0	1
13	0	0	1	1	1	2	1	1
14	0	1	1	1	1	2	0	1
15	1	2	0	1	1	1	0	1
16	0	0	2	2	0	0	2	1
17	1	2	2	2	0	1	0	0
18	0	0	1	2	2	2	1	2
19	0	0	0	2	1	2	0	1
20	1	2	0	2	0	1	1	1
21	1	0	0	1	0	1	1	1
22	0	0	1	2	0	2	0	2
23	2	2	2	2	1	2	0	1
24	1	1	2	2	0	1	1	2
25	0	1	0	1	0	1	1	1
26	1	2	2	2	0	0	1	2
27	0	1	1	1	0	1	2	2
28	1	2	1	2	1	2	1	1
29	0	1	0	1				
30	2	2	1	2				
31	1	2	1	1				
32	1	1	2	2				
Average	0.6250	1.0625	0.7813	1.4688	0.5714	1.3571	0.7857	1.2500

Based on data given in Table 5.8.1, we may apply formulae (5.8.3) and (5.8.4) to obtain estimates, standard errors and their corresponding 95% approximate confidence intervals for τ , β , γ_1 , and γ_2 . The results are summarized in Table 5.8.2. As it can be seen from Table 5.8.2, based on the p-values, the placebo-treatment effect τ is significant at the 0.05 level. The three-month home treatment effect β is also very significant. However, the interactions γ_1 and γ_2 are not significant.

Under the placebo-challenging design, as described above, the group sizes n_1 and n_2 are random and are generally unequal. As the variances of estimators have minimum values when $n_1 = n_2$, a different way of randomization for assigning patients to four groups may be useful. In other words, we may assign patients randomly to four groups, PP, AA, PA, and AP, each consisting of $n/2$ patients, provided that n is even at the first placebo-challenging before the active treatment. In this case, the model and formulae for estimators

Table 5.8.2 Analysis Using Data in Table 5.8.1

Parameter	τ	β	γ_1	γ_2
Estimate	0.1429	0.2969	0.1250	-0.0313
Standard error	0.0717	0.0358	0.0821	0.0717
97.5% lower bound	0.0024	0.2267	-0.0359	-0.1718
97.5% upper bound	0.2834	0.3671	0.2859	0.1092
Two-sided p-value	0.0466	0.0000	0.1286	0.6600

and confidence intervals remain the same, with n_1 and n_2 replaced by $n/2$. However, the efficiency of this design is higher than the one with random n_1 and n_2 , especially when n is small.

5.9 BLINDED READER DESIGNS

Medical imaging drug products are administrated *in vivo* for diagnosis of diseases or monitoring the status of disease conditions. Medical imaging drug products in general are used in conjunction with medical imaging techniques such as radiography, computed tomography (CT), ultrasonography, magnetic resonance imaging (MRI), or radionuclide imaging. There are two general classes of medical imaging drug products: contrast agents and diagnostic radiopharmaceuticals. By increasing the relative difference of imaging signal intensities, contrast agents improve the visualization of tissues, organs, and physiologic processes. Contrast agents can in turn be classified into three categories: (1) iodinated compounds used in radiography and CT, (2) paramagnetic metallic ions used in MRI, and (3) microbubbles, microaerosomes, and other related microparticles used in diagnostic ultrasonography. Diagnostic radiopharmaceutical is either (1) an article that is intended for use in the diagnosis or monitoring of a disease or disease status and that exhibits spontaneous disintegration of unstable nuclei with the emission of nuclear particles or photons or (2) any nonradioactive reagent kit or nuclide generator that is intended to be used in the preparation of such an article. Clinical trials for evaluation of medical imaging drug products should also follow the general principles in design and analysis for clinical trials such as use of a control group, blinding, and randomization of treatments. However, many medical imaging drug products possess unique characteristics that require special considerations in image evaluations for their efficacy and safety. The U.S. FDA recently issued a draft guidance on *Developing Medical Imaging Drugs and Biological Products* to address special issues and challenges in evaluation of medical imaging drug products (FDA, 2003a).

If an investigational medical imaging agent is being developed for an indication for which other drug products or diagnostic modalities have been approved, a direct, concurrent comparison to the comparator should be performed. In addition, the results of evaluation of both investigational imaging products and comparators should be compared not only to one another, but also to an independent gold standard. A *gold standard (truth standard)* is an independent method of measuring the same variable being measured by the investigational imaging drug product that is known or believed to give the true value of the measurement. Although blinding is one of the fundamental principles to minimize any potential bias that could arise in a clinical trial, it is infeasible or even impossible to blind the investigators who administer the investigational medical imaging agents. On the other hand, effectiveness of medical imaging drug products should be evaluated based on the

images by *readers* (usually trained radiologists) obtained with the investigational agents or controls under different conditions or at different times with respect to agent administration. Therefore, despite the fact that administration of medical imaging agents by investigator is unblinded, evaluation of the images by readers can be performed in a blinded fashion that is referred to as *blinded imaging evaluations*. A clinical trial for evaluation of medical imaging products with blinded imaging evaluation is said to have a *blinded reader design*. In addition, for medical imaging drug products, the data and information generated by the blinded reader design are considered as the substantial evidence from the adequate well-controlled trials.

Similar to the various degree of traditional blinding used in clinical trials, blinded imaging evaluations can be also classified into four types: fully blinded image evaluation, image evaluation blinded to outcome, sequential unblinding, and unblinded imaging evaluation (FDA, 2003). For a *fully blinded image evaluation*, readers are blinded to the following information:

1. Results of evaluation of gold standard (or truth standard), of the final diagnosis, of patient outcome.
2. Any patient-specific information, including inclusion/exclusion criteria, details of the protocol, anatomic orientation to the images, history, physical examinations, laboratory results, results of other image studies.
3. Treatment identity.

Unlike the usual definition of blinding, readers in a fully blinded image evaluation are not only masked to the treatments to subjects are assigned to, but also withheld any patient-specific information.

On the other hand, readers in an *image evaluation blinded to outcome* may have knowledge of some particular elements of patient-specific information in (2). If the magnitude of clinical information is given to readers incrementally in successive reading of the same images, this type of image evaluation is referred to as *sequential unblinding image evaluation* that typically is a three-step process:

1. A fully blinded image evaluation is performed. The data of this evaluation should be locked in a secure fashion that it is impossible to change or alter the evaluation when additional clinical information becomes available.
2. An image evaluation blinded to outcome is performed. The data of this evaluation are also locked in the similar manner.
3. The results of the above two types of blinded evaluations are compared to that with the gold standard (or of the final diagnosis, or of patient outcome) to determine diagnostic performance of the medical image drug products.

In an *unblinded image evaluation*, readers are aware of the results of patient evaluation with the gold standard, of patient-specific information, and of treatment identity. Additional patient information provided to readers for an unblinded image evaluation may alter readers' diagnostic assessments and introduce confounding effects and bias into the image evaluations. As a result, the FDA draft guidance suggests that only a fully blinded image evaluation or an image evaluation blinded to outcome serve as the *primary image evaluation* for demonstration of efficacy to support licensing of medical imaging drug products. In addition, they should be conducted through sequential unblinding.

The FDA draft guidance also suggests that at least two blinded readers (and preferably three or more) evaluate images for each trial that is intended to demonstrate efficacy. Image evaluations can be classified by the manner that readers appraise the images and where the images are evaluated. *Independent image evaluations* are those by independent readers that are completely unaware of findings of other readers and are not influenced by the findings of other readers. In addition, the results of each reader's evaluation of images should be locked in a database shortly after it is obtained and before other types of image evaluations are performed so that blinded readers can evaluate images independently. As a result, only the independent image evaluation by blinded readers can serve as the primary image evaluation to establish efficacy of the investigational medical imaging agent. Sometimes different readers evaluate the same set of images together. This type of image evaluation is referred to as *consensus image evaluation*. Although it involves more than one reader, it actually only evaluates a single image and does not fulfill the requirement of the image evaluation by multiple blinded readers mentioned above. For consensus image evaluations, readers do not assess images independently and therefore they cannot serve as the primary image evaluation used to demonstrate efficacy of medical imaging drug products.

To control known or unknown factors that can compromise the integrity and introduce bias to the blinded image evaluations and to ensure that blinded readers conduct their image evaluation independently of other image evaluations, offsite image evaluations are recommended by the draft FDA guidance for the trials that are intended to demonstrate efficacy. *Offsite image evaluations* are image evaluations performed at sites that have not been involved in the conduct of the trial, and by readers who have not had contact with patients, investigators, and other personnel involved in the trial. On the other hand, *onsite image evaluations* are performed by investigators involved with the study or in the care of patients or at sites involved with the conduct of the study. Because readers conducting onsite image evaluations are not blinded to patient-specific information, results of evaluation of gold standard or patient's treatment assignment, onsite image evaluations in general cannot be used for the primary image evaluation to establish efficacy.

One of the key issues in blinded reader designs is the arrangement that different images from the same patient obtained under different conditions or at different time points during the trial that are evaluated by a reader. A *separate image evaluation* is an image evaluation where a reader evaluates a test image obtained from a patient independently of other test images obtained from the same patient. In other words, in a separate image evaluation, readers assess each test image of a patient based on their own merit without reference to or recalls of, any other test images obtained from the same patient. It follows that in a separate image evaluation, readers should not be influenced by the results of evaluations of test images obtained from the same patients. Because a separate image evaluation assesses each individual image either of the same patient or of a different patient independently, it is also referred to as an *unpaired image evaluation* by the draft FDA guidance. For a separate image evaluation, images obtained under different conditions or at different times are first mixed together into a merged set and a sequential identification number is then given to each test image. Random codes are generated to determine the order of image evaluations that the blinded reader should follow. A separate image evaluation provides a mechanism such that multiple test images are not evaluated simultaneously, and the test images are not evaluated sequentially within the same patients. A stratified randomization can be also employed to evaluate test images obtained under different conditions. Here conditions can be considered as stratification factors. A set of separate randomization codes for the order of image evaluations can be generated for each condition. Test images obtained under one

condition then can be assessed individually in an order determined by the random codes, followed by an evaluation of the test images obtained under a different condition using the random codes for that condition. Of course, the order of evaluation for different conditions can also be randomly determined.

Sometimes, a reader may simultaneously evaluate two or more test images obtained from the same patient under different conditions or at different times with respect to administration of medical imaging drug products. This type of simultaneous image evaluations is referred to as *combined image evaluations* or *paired image evaluations*. For example, for contrast agents, both unenhanced and enhanced images may be concurrently evaluated in a comparative manner. It should be noted that combined image evaluations will increase the likelihood of introducing bias to image evaluations. For example, a simultaneous evaluation of images obtained with two different medical imaging drugs for detection of masses may give a biased estimate of difference on diagnostic performance between two drugs. A blinded reader who easily identifies a mass on an image obtained from one drug product might be more likely to identify a mass on a juxtaposed image obtained from the other drug product even though that mass is not clearly seen on the latter image. This phenomenon is called over-reading the presence of mass in a paired comparison. Conversely, under-reading the image is another possible bias that may be introduced by combined image evaluations. To reduce the bias caused by over-reading or under-reading the images, it is suggested that not only order for the set of test images from different patients be evaluated randomly but also simultaneous side-by-side evaluation of images from the same patients be avoided and the order for evaluation of test images from the same patient be also randomly assigned. However, this procedure for combined image evaluations cannot completely eliminate the possible recall bias. Because of inability to eliminate the bias introduced by combined image evaluations, when it is performed, an additional independent separate image evaluation should be completed on at least one of the members of the combination (FDA, 2003a).

As mentioned before, the gold standard provides an independent means of evaluating the same variable being assessed by both the investigational medical imaging drug products and its comparator. Therefore, the gold standard is crucial to establish that the results obtained with the medical imaging drug product are valid and reliable. The draft FDA guidance lists the following principles that should be incorporated prospectively into the design, conduct, and analysis of a clinical trial using a blinded reader design for assessment of medical imaging drug products:

1. The true state of the subjects should be determined with a gold standard without knowledge of the test results obtained from the medical imaging drug product.
2. Test results obtained from the medical imaging drug product should be evaluated without the knowledge of the results obtained from the gold standard.
3. Gold standards should not include as a component of any test results obtained from the medical imaging product.
4. Evaluation of the gold standard should be planned for all enrolled subjects, and the decision to evaluate a subject with the gold standard should not be affected by the test results of the medical imaging drug product under study.

In general, regardless of the types of image evaluations, they should consist of blinded, randomized, independent readings that are designed to demonstrate the efficacy of the investigational medical imaging drug products. Case report forms (CRF) also play an independent

role in reducing the potential bias. For example, to reduce recall bias, different pages in the CRF should be used for two types of image evaluations and each image evaluation should be performed with sufficient time between readings to decrease recall and without reference to prior results. If one of the objectives is to estimate the differences among different types of image evaluations, CRF should contain items or questions that are identical so that differences can be calculated. Evaluation of medical imaging drug products involve not only the design and analysis of the trial but also the design and analysis of blinded reader designs. Randomization and blinding of the treatments as well as randomization and blinding of image evaluations should be taken into account for assessment of efficacy of medical image drug products.

5.10 DISCUSSION

In this chapter we discussed several basic statistical designs, the parallel design, the crossover design, the titration design, and the enrichment design, all of which are commonly employed in clinical trials at various stages of clinical development. Each design has its own merits and limitations under different circumstances. How to select an appropriate design when planning a clinical trial is an important question. The answer to this question depends on many factors, namely those summarized below:

1. Number of treatments to be compared.
2. Characteristics of the treatment.
3. Study objective(s).
4. Availability of patients.
5. Inter- and intrapatient variabilities.
6. Duration of the study.
7. Dropout rates.

For example, when choosing a design from a parallel design and a crossover design, if the intrapatient variability is the same as or larger than the interpatient variability, the inference on the difference in treatments will be the same regardless of which design is used. Actually, a crossover design in this situation would be a poor choice, since blocking results in the loss of some degrees of freedom and will actually lead to a wider confidence interval on the difference between treatments. If a clinical study compares more than three treatments, a crossover design may not be appropriate. The reasons are (1) it may be too time-consuming to complete the study, since a washout period is required between treatment periods, (2) it may not be desirable to switch medications too frequently for each subject due to medical concerns, (3) too many treatment periods may increase the number of dropouts, and (4) the disease status may change from treatment period to treatment period. In this case a balanced incomplete block design is preferred. However, if we compare several test treatments with a placebo control, the within-patient comparison may not be reliable, since the patients in some sequences do not receive the placebo control. If the drug has a very long half-life, and/or it possesses a potential toxicity, or there are carryover effects, then a parallel group design may be a possible choice. With this design the study avoids a possible cumulative toxicity due to the carryover effects from one treatment period to the next. In addition, the study can be completed in less time compared to that of

a crossover design. However, the drawback is that the comparison is made based on the interpatient variability. If the interpatient variability is large relative to the intrapatient variability, the statistical inference on the difference between treatments is not reliable. Even if the interpatient variability is relatively small, a parallel group design may still require more patients in order to reach the same degree of precision achieved by a crossover design. In practice, a crossover design, which can remove the interpatient variability from the comparison between treatments, is often considered to be the design of choice if the number of treatments to be compared is small, say no more than three. If the drug has a very short half-life (i.e., there may not be carryover effects if the length of washout is long enough to eliminate the residual effects), a crossover design may be useful for the assessment of the intrapatient variability provided that the cost for adding one period is comparable to that of adding a patient. In summary, choosing an appropriate design for a clinical trial is an important issue in the development of a study protocol. The selected design may affect the data analysis and the interpretation of the results. Thus all the factors listed above should be carefully evaluated before an appropriate design is chosen.

It, however, should be noted that one of the primary assumptions for a crossover design is that the disease condition remain stable during the study. In practice, this assumption is usually not met. As a result one of the major disadvantages is that spontaneous changes in the disease condition may occur during the study. In this case, although we may establish baseline at each treatment period to eliminate the residual effect, the treatment effect may be confounded with the residual effect. Therefore a crossover design may not be feasible when there are carryover effects. Although a parallel group design is not capable of identifying and removing the interpatient variability from the comparison between treatments, due to its simplicity and easy implementation, it is probably the most commonly used design in clinical phase II and III studies.

One of the controversial issues in clinical trials is the so-called *unethical use* of placebo concurrent control (Rothman and Michels, 1994). To meet this challenge, various variations of placebo-controlled trials have been proposed to minimize the possibility of the use of concurrent placebo control in clinical trials. These include add-on design, replacement design, and randomized withdrawal designs (ICH E10, 1999). An *add-on design* is a placebo-controlled trial of a new drug conducted in patients also receiving standard therapy. This design is useful only when standard therapy is not fully effective and can be used to demonstrate that the new drug can provide additional evidence of improving clinical outcomes. The add-on design is often employed in evaluation of new agents in the treatment of cancer, epilepsy, or heart failure. Because it is an add-on design, the efficacy can be established for the combination therapy. This design, however, is likely to be successful if the new drug uses a pharmacologic mechanism different from that of the standard therapy.

A variation of the add-on design is the *replacement design* in which the new drug or placebo is added by random assignment to the conventional treatment given at an effective dose, and conventional treatment is then withdrawn gradually, usually by tapering. The objective of using a replacement design is to compare the ability to maintain the patient's baseline status between the new drug and placebo. For example, this design has been used to investigate steroid-sparing substitution in steroid-dependent patients without need for initial steroid withdrawal and recrudescence of symptoms in a washout period.

The *randomized withdrawal design* usually consists of two phases. For a trial using the randomized withdrawal design, patients receive an investigational drug in the first phase of a prespecified length and are then randomly assigned to continue to receive the investigational drug or placebo in the second phase. The active treatment is actually withdrawn for

the patients receiving placebo in the second period. The first phase of the randomized withdrawal design is a prerandomization observation period to establish the initial on-therapy baseline on the investigational drug and in general is longer than the second period. The objective of this phase is to investigate the long-term persistence of effectiveness of the drug when long-term use of placebo is not acceptable. The second phase is a withdrawal phase and a post-randomization observation period. Any differences that are observed between groups receiving continued drug and placebo in this period would demonstrate the effect of the investigational drug. This period can also use early escape or time-to-event endpoints to minimize the exposure of subjects to placebo. As a result, one of the major advantages of the randomized withdrawal design, when used jointly with an early escape endpoint, is that the period of placebo exposure with poor response that a patient would have to undergo is short. In addition, the randomized withdrawal design can also be used to investigate the dose-response relationship of the investigational drug. After all patients receive an initial fixed dose, they are randomly assigned to several doses and placebo in the withdrawal phase. If the first phase of the randomized withdrawal design is a placebo-controlled titration design, it is an enrichment design with responders randomly assigned to receive several doses and placebo in the withdrawal phase. The joint utilization of titration design and randomized withdrawal design enables us to investigate dose-response rigorously while allowing the efficiency of the titration design.

6

DESIGNS FOR CANCER CLINICAL TRIALS

6.1 INTRODUCTION

When designing cancer clinical trials for development and evaluation of therapeutic interventions, two special aspects must be taken into consideration. The first aspect is the target patient population. In practice, patients in cancer clinical trials are those with malignant tumors. Unlike other diseases, most cancers are life-threatening diseases in which the disease process is usually irreversible, and in most cases, they are neither curable nor controllable. In addition, patients with malignant tumors have limited life expectancy. The other issue is that most of the anti-cancer drugs under investigation are cytotoxic agents that usually have a very narrow therapeutic window with the following dilemma. At the lower dose, these cytotoxic agents provide little or no efficacy but can generate more severe and irreversible toxicities than most of pharmaceutical agents for treatment of other diseases such as immunosuppression, hepatic, renal, or cardiac toxicity. Effectiveness of these anti-cancer cytotoxic agents can be delivered only at higher doses that may also induce fatal or life-threatening serious adverse events. To resolve this dilemma, we may select an appropriate study design for evaluation of the cancer therapeutic interventions for phase I and II trials based on the following criteria that:

1. It minimizes exposure of subjects to these therapeutic interventions.
2. It selects efficacious cytotoxic agents with an acceptable safety profile in the most efficient manner.

As mentioned above, therapeutic agents for cancer treatment can induce severe safety concern even at lower dose levels. As a result, unlike employing healthy normal volunteers for

phase I safety and tolerance studies of pharmaceutical agents, phase I trials for new anti-cancer agents are often conducted on terminal cancer patients for which the test cytotoxic drugs may be the last hope. The primary scientific objective of the evaluation of new chemotherapeutic agents in cancer patients during the phase I clinical development is to employ an efficient, reliable, but yet practical dose-finding design to search the maximum dose with an acceptable and manageable safety profile for use in subsequent phase II trials. This dose with an acceptable and manageable safety profile is usually referred to as the maximum tolerable dose (MTD). The unacceptable or unmanageable safety profile is in general called the dose-limiting toxicity (DLT), which is predefined by some criteria such as Grade 3 or greater hematological toxicity according to the United States National Cancer Institute's Common Toxicity Criteria (CTC). In summary, the MTD is the highest possible but still tolerable dose with respect to some prespecified dose-limiting toxicity (see, e.g., Storer, 1993; Korn et al., 1994).

For most cancer treatments, the drug must be delivered at the maximum dose for achieving the maximum effect. Therefore, once the MTD for a new anti-cancer agent is determined from phase I trials, then the anti-tumor activity of the drug can be evaluated at the MTD during the phase II clinical development for which trials usually consist of a single treatment without a comparative or controlled group. The objective of phase II cancer trials is to quickly determine whether the new anti-cancer agent has sufficient activity against a particular type of tumor to justify its further development. As a result, the objective of a phase II cancer trial can be translated in the following statistical hypothesis:

H_0 : The anti-tumor activity is below some undesirable level, say p_0 .

vs. H_a : The anti-tumor activity is above or equal to some targeted level, say p_1 .

(6.1.1)

The reason for using the anti-tumor activity as the primary endpoint for phase II cancer trials is that it can be observed in a considerably shorter period of time than the usual survival endpoint used for more rigorous phase III trials. The anti-tumor activity is usually measured by the degree of tumor shrinkage by various different criteria that are often referred to as the objective tumor response. The World Health Organization's (WHO) definition of the objective tumor response (WHO, 1979) and the one suggested by Miller et al. (1981) are probably the earliest two criteria for evaluation of anti-tumor activity for a cytotoxic agent. These objective tumor responses are based on the 2D measurements of tumors. In 1994, the European Organization for Research and Treatment of Cancer (EROTC), the National Cancer Institutes of the United States, and the National Cancer Institute of Canada Clinical Trial Group set up a task force to review the above criteria for evaluation of the response to treatment in solid tumors. In 2000, the task force published a guideline that proposed a new method called *Response Evaluation Criteria in Solid Tumors (RECIST)* based on 1D tumor measurement to evaluate the anti-tumor activity (Therasse et al., 2000).

In the next section, some general considerations for phase I cancer clinical trials are addressed. In Section 6.3, we introduce single-stage up-and-down designs, including the standard dose escalation design for phase I cancer trials for determination of MTD. To overcome the shortcomings of the standard dose escalation design, two-stage up-and-down designs, including the accelerated titration phase I design proposed by Simon et al. (1997) are reviewed in Section 6.4. In addition, in Section 6.5, Bayesian approaches such as continual reassessment method (CRM) and its variations are discussed. In Section 6.6, we introduce optimal and flexible multiple-stage designs, including the commonly used Simon

optimal two-stage design for phase II cancer studies. Section 6.7 reviews the randomized phase II design proposed by Simon et al. (1985). Final remarks and discussion are given in Section 6.8.

6.2 GENERAL CONSIDERATIONS FOR PHASE I CANCER CLINICAL TRIALS

In practice, MTD could be statistically interpreted as some percentile of a tolerance distribution or dose-response curve in terms of the presence or absence of DLT. In other words, the MTD is the dose where a specified proportion of patients, say, p_0 , experience DLT. Storer (1997) indicated that the value of p_0 is usually in the range from 0.1 to 0.4. Let Y be the binary response such that $Y = 1$ denote the occurrence of a predefined DLT and $\{d_i, i = 1, 2, \dots, I\}$ be a set of fixed dose levels. Relationship between the occurrence of the DLT and dose level d can be described by the following model:

$$\text{logit}[P(x, \theta)] = \alpha + \beta x, \quad (6.2.1)$$

where $\text{logit}(\cdot)$ is the logistic function of the probability of the occurrence of the DLT, x is the dose level taking one of the values d_i , and $\theta = (\alpha, \beta)'$. Thus, the MTD is then defined as

$$x_m = (k_p - \alpha)/\beta,$$

where $k_p = \text{logit}(p_0) = \ln[p_0/(1 - p_0)]$ and \ln denotes the natural logarithm. On the other hand, the dose levels employed in the phase I cancer trials for determination of the MTD are generally derived from the information of animal studies. Storer (1997) pointed out that a commonly employed starting dose level is from one-tenth to one-third of the mouse LD₁₀. In addition, the dose levels are usually selected to be approximately equally spaced on the logarithmic scale. Schneiderman (1967), for example, suggested the use of the modified Fibonacci sequence of the diminishing multipliers of {2, 1.67, 1.5, 1.4, 1.33, ...}.

As indicated earlier, the main purpose of phase I cancer trials is to establish the MTD with an adequate precision. The following considerations are important for selection of an appropriate design in phase I trials for estimation of the MTD:

1. The patients are critically ill. Some of them are even in the terminal stage of the disease and the test anti-cancer agent may be the last hope for the patients.
2. The number of patients available for phase I cancer trials is relatively small.
3. The patient population is usually rather heterogeneous because phase I cancer trials might enroll terminal cancer patients with different types of malignant tumor at various disease stages.
4. Phase I cancer trials can be viewed as a screening process where anti-cancer cytotoxic agents with a tolerable safety profile are selected and their MTDs are determined with a minimal number of patients in a minimal amount of time.
5. Most anti-cancer agents generally can induce serious, irreversible, life-threatening, or even fatal toxicity. Thus, phase I cancer trials are usually conducted to establish

the MTD from below. In fact, regulatory agencies sometimes dictate the dose for the first patient.

Designs for phase I cancer trials generally can be classified into three categories: single-stage design, two-stage design, and Bayesian design. The single-stage and two-stage designs are up-and-down designs where the doses are adjusted either upward (escalation) or downward (de-escalation) within the prespecified set of the fixed dose levels. The Bayesian approach chooses the dose level for the next patients by minimizing some measure based on the difference between the current estimate of probability for the occurrence of DLT and p_0 .

6.3 SINGLE-STAGE UP-AND-DOWN PHASE I DESIGNS

As indicated in Storer (1989, 1993, 2001), three single-stage designs, namely, design A, design B, and design D are commonly employed in phase I cancer trials. In what follows, we provide a brief description for each of these designs.

6.3.1 Design A—Standard Dose Escalation Design

For design A, we start with a group of three patients, who are treated at the lowest dose level. At the second step, if no prespecified DLT is observed in all three patients, then the dose for the next group of three patients is escalated to the next higher dose level. Otherwise, the next group of three patients is treated at the same dose level. In Step 3, the dose of the next group of three patients is escalated to the next higher dose level if the prespecified DLT is observed at most in one patient of the six patients from both Steps 1 and 2, otherwise, the trial stops. For Step 4, we repeat Steps 2 and 3 with two consecutive groups of three patients until the trial stops. A flowchart for design A is given in Figure 6.3.1. Traditionally, if the study stops at dose level, d_i , then the MTD is estimated as the next lower dose level, d_{i-1} . If the trial stops at the dose level where only three patients were treated, then an additional three patients should be enrolled at the next lower dose level for a total of six patients. As a result, the MTD is the highest dose where at most one-sixth of the patients developed DLT.

It can be seen that the standard dose escalation design only allows dose level to be escalated upward. As a consequence, the starting dose level is the lowest dose level d_1 , and hence, lots of patients are treated at the doses well below the therapeutically meaningful level. Therefore, the MTD obtained from this design may be too conservative. Another drawback of this design is that because too many patients are treated at the lower dose levels, it might take a long time for the MTD to be reached. To overcome these two drawbacks, Storer (1989) proposed the following three designs, which not only allow both escalation and de-escalation, but also do not require the use of the lowest dose level as the starting dose.

6.3.2 Design B

For design B, we start with a single patient at a preselected dose level. For the next step, if no prespecified DLT is observed in this patient, then the next patient is treated at the same dose level; otherwise, the next patient is treated with the next lower dose level. At the third

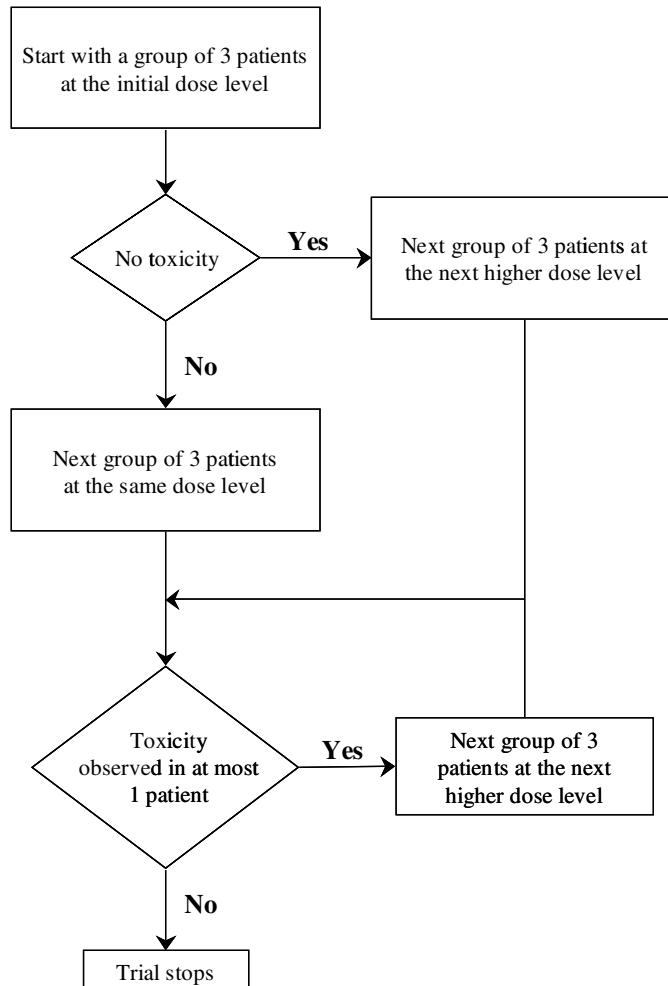
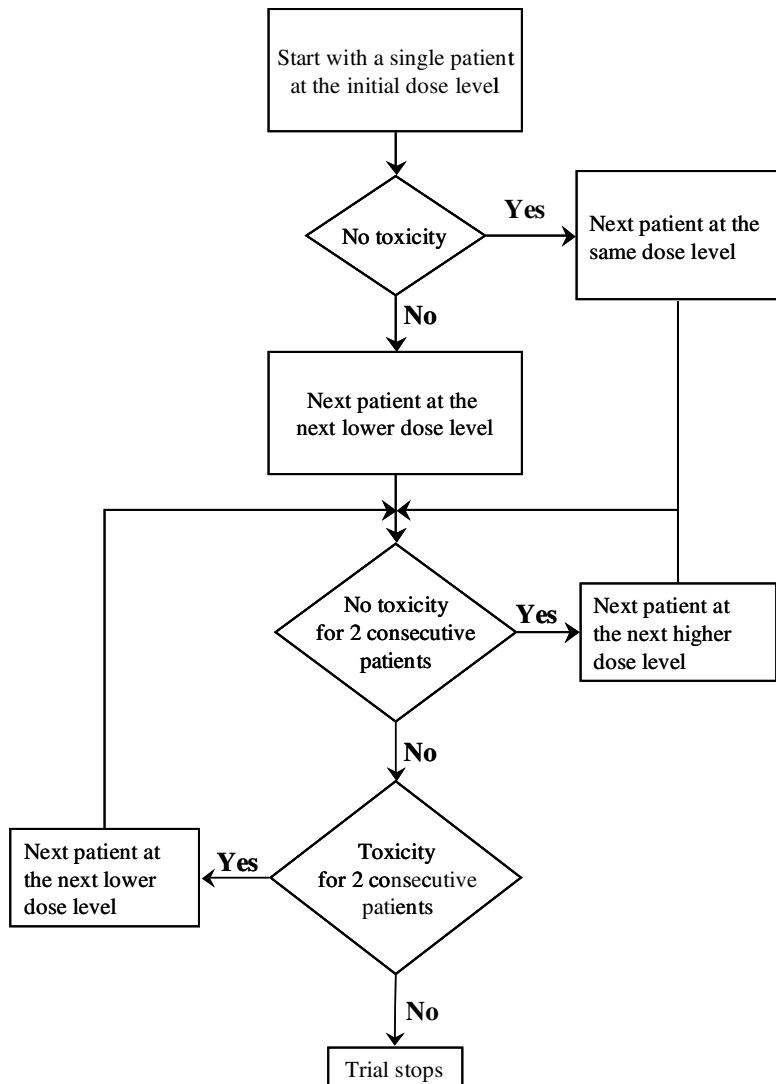


Figure 6.3.1 Flowchart for design A.

step, if no prespecified DLT occurs in both consecutive patients, then the next patient is treated with the next higher dose level. If the DLT is observed in both consecutive patients, then the dose of the next patient is deescalated to the next lower dose level. Otherwise, the trial stops. In Step 4, we repeat Steps 2 and 3 until the trial stops. A flowchart for design B is given in Figure 6.3.2.

6.3.3 Design D

For design D, similar to design A, we start with a group of three patients who are treated at an initial dose d_i , $i = 1, 2, \dots, I$. In the next step, the dose of the next group of three patients is escalated to the next higher dose level if no DLT is observed in all three patients, or stays at the same dose level if the prespecified DLT is observed in one patient, or deescalates to

**Figure 6.3.2** Flowchart for design B.

the next lower level if the DLT occurs in more than one patient. At Step 3, we continue Step 2 until all of the prespecified number of the patients have completed the study. A flowchart for design D is given in Figure 6.3.3.

6.4 TWO-STAGE UP-AND-DOWN PHASE I DESIGNS

In a simulation study, Storer (1989) reported that none of the single-stage designs performs well in an arbitrary dose-response setting with a fixed sample size. In addition, the standard

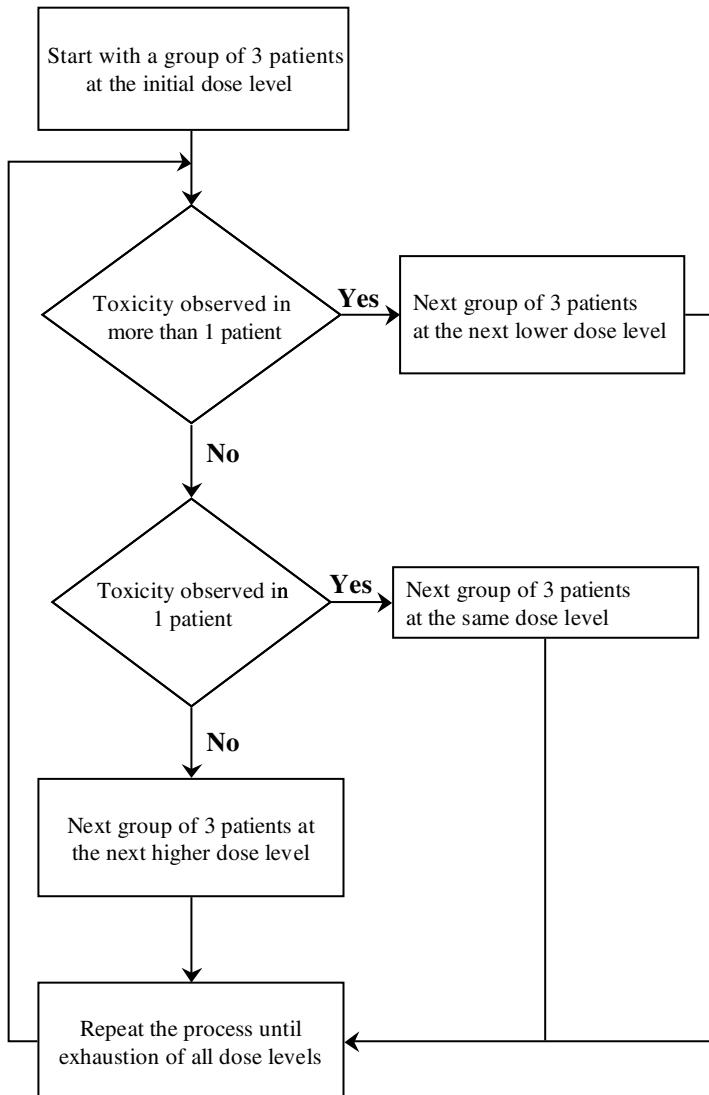


Figure 6.3.3 Flowchart for design D.

dose escalation design frequently fails to provide a convergent estimate of MTD. As a result, Storer (1989, 1993) and Simon et al. (1997) proposed a variety of combinations of traditional single-stage designs to overcome the drawbacks of the single-stage phase I designs. The ideas are to:

1. Concentrate sampling around the MTD.
2. Reduce the duration of the study.
3. Minimize the number of patients treated at subtherapeutic dose levels.
4. Obtain the information about inter-patient variability and cumulative toxicity.

6.4.1 Design BD (Storer, 1989, 2001)

Storer (1989) proposed a two-stage design that combines design B and design D. We will denote this design by design BD.

For design BD, we start design B until the trial stops according to the stopping rule described in design B. At Stage 2, we then continue the trial with design D. If a DLT is observed in the last patient at the first stage, then the initial dose for the second stage for design D is the next lower dose level with respect to the dose of the last patient at the first stage. However, if the last patient at the first stage does not exhibit any prespecified DLT, then the beginning dose for the second stage is the dose of the last patient at the first stage. For all designs described above, the prespecified DLT is observed during the first course of treatment and there is no titration within each individual patient. In other words, there is no intrapatient dose modification for these four designs. Except for the standard dose escalation design, the MTD under designs B, D, and BD needs to be estimated based on the data by using a formal statistical inference procedure.

6.4.2 Accelerated Titration Designs

For the accelerated titration designs (ATD), Simon et al. (1997) further classified the toxicity grades based on the NCI's CTC into two categories: moderate toxicity for grade 0–2 or DLT for grade 3 and above. In addition, the patients will receive at least three courses of treatment in the accelerated titration designs that consist of an initial accelerated stage and a standard dose escalation stage. In other words, there is a dose titration within individual patients. Three types of ATD are described below:

ATD1

For the initial accelerated stage, cohorts of one new patient per dose level start at the lowest dose level. If the first instance of the DLT is observed at the first course in one patient or the two patients exhibit grade 2 toxicity of any type during the first course of treatment, then dose escalation stops and reverts to the standard dose escalation design. For the standard dose escalation stage, the cohort of current dose level is expanded to three patients and continues the trial using the standard dose escalation design in a cohort of three patients. The intrapatient dose escalation is allowed if the worst toxicity is grade 0–1 in the previous course for that patient. The dose is deescalated if grade 3 or above toxicity occurs in the previous course. Otherwise, patients will stay at the same dose level. Dose escalation of the ATD1 uses 40% dose-step increments.

ATD2

Same as ATD1 except that 100% dose steps are used for the initial accelerated stage.

ATD3

Same as ATD2 except that the switch to standard dose escalation design when the first instance of the DLT in any course or the second instance of any course grade 2 toxicity of any type is observed.

Both design BD and various versions of ATD provide the possibility of speeding up the trial and reducing the number of patients assigned to lower dose levels. They use the first instance of DLT observed during the first course of treatment to trigger the switch to the

Table 6.4.1 Summary of Simulations Results of Comparison between Standard Dose Escalation Design and Accelerated Titration Design

Design	Sample Size	Average Number of Patients		
		Grade 1	Grade 2	Grade 3
A	39.9	23.3	5.5	1.9
ATD1	24.4	7.9	6.2	3.0
ATD2	20.7	3.9	6.8	4.3
ATD3	21.2	4.8	6.2	3.2

Summarized from Simon et al. (1997).

traditional standard dose escalation design. However, ATD proposed by Simon et al. (1997) also use the grade 2 toxicity observed also during the first course to provide additional caution. ATD2 and ATD3 allow more rapid dose escalation than ATD1 by using double-dose step during the initial accelerated stage. However, these more aggressive dose escalation schemes of ATDs may be associated with more risk. To investigate the likelihood of increasing risk, Simon et al. (1997) performed an extensive simulation study in which the 1000 sets of simulated data were generated from the parameters estimated from each of 20 different actual phase I trials of 9 different drugs. The results are summarized in Table 6.4.1.

From Table 6.4.1, it is obvious that the sample size required for the accelerated titration designs reduces by at least 40% as compared to the standard dose escalation design. As a result, the duration of ATD trials can be also shortened considerably. In addition, the risk associated with ATD1 and ATD3 appear acceptable. Because the ATDs employ a dose titration scheme within each patient, more information is generated for estimation of the population distribution of MTD, the degree of cumulative toxicity, and the intrapatient and interpatient variability. Simon et al. (1997) indicated that if interpatient variability is small, a fixed dose-regimen can be used in phase II cancer trials, and fewer patients will be either overdosed or underdosed. On the other hand, the use of ATDs requires careful patient management to track the toxicity over multiple courses and clear definitions for DLT and toxicity level considered low that intrapatient titration is acceptable.

6.5 CONTINUAL REASSESSMENT METHOD PHASE I DESIGNS

As mentioned above, patients in cancer phase I trials are often those with terminal cancer and at high risk of death. The characteristics of the anti-cancer cytotoxic agents evaluated in phase I cancer trials are that (1) they produce fatal toxicity at a higher dose level, (2) they yield little or no effectiveness at lower dose levels, and (3) except for some scarce animal data, no information about the dosing range is available. The standard dose escalation design and one-stage or two-stage up-and-down designs update the dose for the next patient based on the information of occurrence of DLT from the current and previous patients. These designs, however, fail quantitatively to employ a model to combine the prior information about the MTD and that from the patients collected in the phase I trials. To address these drawbacks, O'Quigley et al. (1990) proposed the continual reassessment method (CRM) that updates the information of the dose-response relationship through

a Bayesian framework as observations on DLT becomes available and then to use this information to concentrate the trial around the dose that might correspond to the anticipated target toxicity level.

Recall that the MTD is the dose associated with the $100p_0$ percentile of the dose response relationship with respect to the occurrence of a predefined DLT. Let p_1, p_2, \dots, p_I be the initial guesses of probabilities of DLT corresponding to the set of fixed dose level d_1, d_2, \dots, d_I , that are assumed to be monotonically increasing with the dose level. The dose response relationship can then be characterized by a logistic regression model

$$\text{logit}[P(x_i, \beta)] = \alpha + \beta x_i,$$

where the intercept α is assumed fixed (see, e.g., Goodman et al., 1995; Ahn, 1998; Ishizuka and Ohashi, 2001), the slope β is to be estimated, and x_i is updated after the information of the occurrence of DLT for the current patient becomes available using $x_i = \text{logit}^{-1}[P(x_i, \beta)]$. The procedure of CRM is then outlined below:

Initially, we choose a dose response model to be employed in the study and select a set of fixed dose levels with their corresponding probabilities of DLT. Then, choose a fixed sample size n (usually 18 to 24) for the study. For the next step, a prior distribution of the slope, denoted by $g(\beta)$, that reflects the current brief of the investigator about the dose response relationship is chosen. The dose for the first patient is determined as the dose level that produces the prior probability of DLT closest to the targeted probability p_0 . After the result of the occurrence of DLT for the current patient at dose level x_{i-1} becomes available, obtain the Bayesian estimate of slope β with respect to quadratic error loss function as

$$\beta_i = E(\beta|\Lambda_{i-1}) = \int \beta f(\beta|\Lambda_{i-1})d\beta,$$

where $f(\beta|\Lambda_{i-1})$ is the posterior density, which is given by

$$f(\beta|\Lambda_{i-1}) = q_{\Lambda_{i-1}}(\beta)g(\beta)/\int q_{\Lambda_{i-1}}(z)g(z)dz,$$

and $q_{\Lambda_{i-1}}(\beta)$ is the likelihood function such that

$$q_{\Lambda_{i-1}}(\beta) = \prod [P(x_j, \beta)]^{y_j} [1 - P(x_j, \beta)]^{1-y_j},$$

and $\Lambda_{i-1} = \{(x_1, y_1), \dots, (x_{i-1}, y_{i-1})\}$, the cumulative data up to the patient receiving dose level x_{i-1} .

The dose level for the next patient is then determined as the dose level that minimizes the absolute loss function $|P(x_i, \beta_i) - p_0|$ with $x_i = \text{logit}^{-1}[P(x_i, \beta_i)]$, $i = 1, 2, \dots, I$. We then repeat the above steps until the result of the last patient is known. Finally, the MTD, x_m , is estimated as the dose level for the hypothetical $n+1$ patient. The corresponding probability for the occurrence of DLT can then be estimated as $P(x_m, \beta_m)$ with the corresponding $100(1-\alpha)$ credibility interval given as $[P(x_m, \beta_L), P(x_m, \beta_U)]$, where (β_L, β_U) is the $100(1-\alpha)$ credibility interval based on the posterior probability of slope β .

The CRM represents a revolutionary approach to estimation of the MTD as compared to the traditional up-and-down methods. It estimates the MTD from a continuous spectrum of doses, whereas the other methods choose their MTD from a discrete set of prespecified dose levels. In addition, the CRM utilizes a statistical model to synthesize the prior belief about the MTD and cumulative information from the patients in the current trial for selection of

the dose level for the next patient. However, despite these advantages, as pointed out by Korn et al. (1994), Goodman et al. (1995), and Ahn (1998), the original CRM is not free of problems in its implementation for estimation of the MTD in phase I cancer clinical trials. First, the dose level for the next patient can be determined only after the result on the DLT for the current patient becomes available. On the other hand, the CRM only uses a cohort size of one patient for the dose adjustment. Consequently, it may take much longer to complete the trial than the traditional dose escalation design and most other up-and-down designs. Secondly, the CRM might start the trial with an initial dose above the lowest dose that is often one-tenth the LD₁₀ in mice. This possibility of the initial dose above the lowest dose level for the first patient makes many clinicians and regulatory agencies uncomfortable and reluctant to implement the CRM. Another special feature of the CRM that also reduces its acceptance is the possibility that dose can be escalated for more than one dose level. This can result in more patients being treated at high dose levels and being exposed to higher risks. Møller (1995) demonstrated that the second dose can be escalated to dose level 10 even when the first patient has no DLT at the lowest dose level 1 as the initial dose.

Goodman et al. (1995), O'Quigley and Shen (1996), Ahn (1998), Heyd and Carlin (1999), Thall et al. (1999), Zohar and Chevret (2001), Storer (2001), and Ishizuka and Ohashi (2001) have suggested various ways to overcome the above-mentioned drawbacks of the CRM. These improvements and modifications can be summarized as below:

1. The trial always starts with the lowest dose level as the initial dose for the first patient.
2. The size of cohorts increases to three patients.
3. The magnitude of dose escalation is limited to one dose level only between two adjacent cohorts.
4. Use of likelihood approach.
5. Employment of different stopping rules.
6. Different accrual strategies with delayed patient outcomes.
7. Add a model-fitting step at the end for estimation of MTD.
8. Use a two-stage CRM. At the first stage, an up-and-down design is employed until the first DLT, and then the CRM is implemented to incorporate all information obtained at that point.

The CRM can be further modified to incorporate other information and to achieve other objectives. Piantadosi and Liu (1996) and Ishizuka and Ohashi (2001) illustrated that additional pharmacokinetic data can be incorporated into the CRM. Legedza and Ibrahim (2000) proposed a longitudinal design for phase I cancer trials based on the CRM to investigate the cumulative toxicity of anti-tumor cytotoxic agents. Kramar et al. (1999) applied CRM to the combination of two drugs. Gasparini and Eisele (2000) proposed the use of the product-of-beta prior for the curve-free application of the CRM. Recently, Thall et al. (2001), through the framework of the CRM, developed a method for dose-finding on feasibility and toxicity in T-cell infusion trials. Piantadosi et al. (1998) and Dougherty et al. (1999) provided excellent illustrations for practical implementation of the CRM with real examples. Computer software for execution of various versions of the CRM is available in public domain through the Internet, for example, http://odin.mdacc.tmc.edu/anonftp/page_2.html#CRM by Section of Computer Science, Department of Biomathematics, University of Texas M.D. Anderson Hospital.

6.6 OPTIMAL/FLEXIBLE MULTIPLE-STAGE DESIGNS

In phase II cancer trials, it is undesirable to stop a study early when the test drug is promising. On the other hand, it is desirable to terminate the study as early as possible when the treatment is not effective. For this purpose, an optimal multiple-stage design is often employed to determine whether a study drug holds sufficient promise to warrant further testing. In what follows, we will first consider phase II cancer trials with single arm. Then, we will consider multiple-stage designs for trials with multiple-arm.

6.6.1 Single-Arm Trials

Optimal multiple-stage designs that are commonly employed in phase II cancer trials with single arm include optimal multiple-stage designs (e.g., minimax design and Simon's optimal two-stage design) and flexible multiple-stage designs; see, e.g., Simon, 1989; Ensign et al., 1994; Chen, 1997; Chen and Ng, 1998; Sargent and Goldberg, 2001.

Optimal Two-Stage Designs The concept of an optimal two-stage design is to permit early stopping when a moderately long sequence of initial failure occurs. Denote by the number of subjects studied in the first and second stage by n_1 and n_2 , respectively. Under a two-stage design, n_1 patients are treated at the first stage. If there are fewer than r_1 responses, then stop the trial. Otherwise, stage 2 is implemented by including the other n_2 patients. A decision regarding whether the test drug is a promising compound is then made based on the response rate of the $N = n_1 + n_2$ subjects. Let p_0 be the undesirable response rate and p_1 be the desirable response rate ($p_1 > p_0$). If the response rate of a test drug is at the undesirable level, one wishes to reject it as an ineffective compound with a high probability (or the false-positive rate is low), and if its response rate is at the desirable level, not to reject it as a promising compound with a high probability (or the false-negative rate is low). As a result, it is of interest to test the following hypotheses:

$$H_0: p \leq p_0 \quad \text{vs.} \quad H_a: p \geq p_1$$

Rejection of H_0 (or H_a) means that further (or not further) study of the test drug should be carried out. Note that under the above hypotheses, the usual type I error is the false-positive rate in accepting an ineffective drug and the type II error is the false-negative rate in rejecting a promising compound.

To select among possible two-stage designs with specific type I and type II errors, Simon (1989) proposed to select the optimal design that achieves the minimum expected sample size when the response rate is p_0 . Let EN be the expected sample size. Then, EN can be obtained as

$$EN = n_1 + (1 - PET)n_2$$

where PET is the probability of early termination after the first stage, which depends on the true probability of response p . At the end of the first stage, we would terminate the trial early and reject the test if r_1 or fewer responses are observed. As a result, PET is given by

$$PET = B(r_1; p, n_1)$$

where B denotes the cumulative binomial distribution. We would reject the test drug at the end of the second stage if r or fewer responses are observed. Hence, the probability of rejecting the test drug with success probability p is given by

$$B(r_1; p, n_1) + \sum_{x=r_1+1}^{\min(n_1, r)} b(x; p, n_1) B(r - x; p, n_2)$$

where b denotes the binomial probability density function. For specified values of p_0 , p_1 , α , and β , Simon's optimal two-stage design can be obtained as the two-stage design that satisfies the error constraints and minimizes the expected sample size when the response probability is p_0 . As an alternative design, Simon (1989) also proposed to seek the minimum total sample size first and then achieve the minimum expected sample size for the fixed total sample size when the response rate is p_0 . This design is referred to as the minimax design.

Example 6.6.1 Suppose a sponsor is interested in conducting a single-arm cancer trial with an optimal two-stage design. p_0 and p_1 were chosen to be 0.20 and 0.40, respectively. Based on the discussion above, the optimal two-stage design gives (3/13, 12/43) for achieving an 80% power at the 5% level of significance. In other words, at the first stage, 13 subjects are tested. If no more than 3 subjects respond, then terminate the trial. Otherwise, accrual continues to a total of 43 subjects. We would conclude that the test drug is effective if there are more than 12 (out of 43 subjects) responses.

Flexible Two-Stage Designs As an alternative to the optimal two-stage designs described above, Chen and Ng (1998) proposed optimal multiple-stage flexible designs for phase II trials by simply assuming that the sample sizes are uniformly distributed on a set of k consecutive possible values. As an example, the procedure for obtaining an optimal two-stage flexible design is outlined below.

Let r_i and n_i be the critical value and the sample size for the first stage and R_j and N_j be the critical value and sample size for the second stage. Thus, for a given combination of (n_i, N_j) , the expected sample size is given by

$$EN = n_i + (1 - PET)(N_j - n_i)$$

where

$$PET = B(r_i; p, n_i)$$

$$= \sum_{x \leq r_i} b(x; p, n_i)$$

The probability of rejecting the test drug for (n_i, N_j) is then given by

$$B(r_i; p, n_i) + \sum_{x=r_i+1}^{\min(n_i, R_j)} b(x; p, n_i) B(R_j - x; p, N_j - n_i)$$

The average probability of an early termination ($APET$) is the average of PET for all possible n_i . The average total probability of rejecting the test drug ($ATPRT$) is the average of the above probability for all possible combinations of (n_i, N_j) . The average expected sample size (AEN) is the average of EN . Chen and Ng (1998) considered the following criteria

for obtaining an optimal flexible design. If the true response rate is p_0 , we reject the test drug with a very high probability (i.e., $ATPRT \geq 1 - \alpha$). If the true response rate is p_1 , we reject the test drug with a very low probability (i.e., $ATPRT \leq \beta$). There are many solutions of (r_i, n_i, R_j, N_j) 's that satisfy the α and β requirements for the specific p_0 and p_1 . The optimal design is the one that has minimum AEN when $p = p_0$. The minimax design is the one that has the minimum N_k and the minimum AEN within this fixed N_k when $p = p_0$.

Example 6.6.2 Consider the same example as described in Example 6.6.1. Now, suppose the sponsor is interested in conducting a single-arm cancer trial with a flexible two-stage design for achieving a 90% power at the 10% level of significance. It is also decided to choose $p_0 = 0.1$ and $p_1 = 0.30$. In this case, the flexible two-stage design gives (1/11–17, 2/18) for the first stage and (3/24, 4/25–28, 5/29–31) for the second stage for achieving a 90% power at the 10% level of significance. The optimal flexible two-stage design allows the first stage sample size to range from 11 (n_1) to 18 (n_8). The rejection boundary r_i is 1 if n_i ranges from 11 to 17, and 2 if n_i is 18. If the observed responses are greater than r_i , we accrue $27 - n_i$ additional subjects at the second stage. The flexible optimal two-stage design allows the total sample size to range from 24 (N_1) to 31 (N_8). The rejection boundary R_j is 3 if N_j is 24, 4 if N_j ranges from 25 to 28, and 5 if N_j ranges from 29 to 31.

Optimal Three-stage Designs The disadvantage of a two-stage design is that it does not allow early termination if there is a long run of failures at the start. To overcome this disadvantage, Ensign et al. (1994) proposed an optimal three-stage design, which modifies the optimal two-stage design. The optimal three-stage design is implemented by testing the following similar hypotheses:

$$H_0: p \leq p_0 \quad \text{vs.} \quad H_a: p \geq p_1$$

Rejection of H_0 (or H_a) means that further (or not further) study of the test drug should be carried out. At stage 1, n_1 patients are treated. We would reject H_a (i.e., the test treatment is not responding) and stop the trial if there is no responses. If there are one or more responses, then proceed to stage 2 by including additional n_2 patients. We would reject H_a and stop the trial if the total number of responses is less than or equal to a prespecified number of r_2 ; otherwise, continue to stage 3. At stage 3, n_3 more patients are treated. We would reject H_a if the total responses for the three stages combined is less than or equal to r_3 . In this case, we conclude that the test drug is ineffective. On the other hand, if there are more than r_3 responses, we reject H_0 and conclude that the test drug is effective. Based on the concept of the above three-stage design, Ensign et al. (1994) considered the following to determine the sample size. For each value of n_1 satisfying

$$(1 - p_1)^{n_1} < \beta$$

where

$$\beta = P(\text{reject } H_a | p_1)$$

computing the values of r_2 , n_2 , r_3 , and n_3 that minimize the null expected sample size $EN(p_0)$ subject to the error constraints α and β , where

$$EN(p) = n_1 + n_2\{1 - \beta_l(p)\} + n_3\{1 - \beta_l(p) - \beta_2(p)\}$$

and β_i are the probability of making type II error evaluated at stage i . Ensign et al. (1994) use the value of

$$\beta = (1 - p_1)^{n_i}$$

as the type II error rate in the optimization along with type I error

$$\alpha = P(\text{reject } H_0 | p_0)$$

to obtain r_2 , n_2 , r_3 , and n_3 . Repeating this, n_i can then be chosen to minimize the overall $EN(p_0)$.

Example 6.6.3 Similar to Example 6.6.1, suppose a sponsor is interested in conducting a single-arm cancer trial with an optimal three-stage design. p_0 and p_1 were chosen to be 0.25 and 0.40, respectively. Based on the discussion above, the optimal three-stage design gives (0/6, 7/26, 24/75) for achieving an 80% power at the 5% level of significance. In other words, at the first stage, six subjects are treated. If no responses are seen, then the trial is terminated. Otherwise, accrual continues to a total of 26 subjects at the second stage. If no more than 7 subjects respond, then stop the trial. Otherwise, proceed to the third stage by recruiting an additional 49 subjects. We would conclude that the test drug is effective if there are more than 24 responses for the subjects in the three stages combined.

Note that the optimal three-stage designs proposed by Esign et al. (1994) restricts the rejection region in the first stage to be zero response, and the sample size to at least 5. As an alternative, Chen (1997) also extended Simon's two-stage to a three-stage design without these restrictions. As a result, sample sizes can be obtained by computing the values of r_1 , n_1 , r_2 , n_2 , r_3 , and n_3 that minimize the expected sample size

$$EN = n_1 + (1 - PET_1)n_2 + (1 - PET_{all})n_3$$

$$PET_1 = B(r_1; n_1, p) = \sum_{x \leq r_1} b(x; n_1, p)$$

$$PET_{all} = PET_1 + \sum_{x=r_1+1}^{\min(n_1, r_2)} b(d; n, p) B(r_2 - x; n_2, p).$$

Example 6.6.4 Suppose the sponsor is interested in conducting a single-arm cancer trial with the optimal three-stage design as described above. With $p_0 = 0.25$ and $p_1 = 0.40$, the optimal three-stage design gives (4/17, 12/42, 25/79) for achieving an 80% power at the 5% level of significance. In other words, at the first stage, 17 subjects are treated. If no more than four responses are seen, then the trial is terminated. Otherwise, accrual continues to a total of 42 subjects at the second stage. If no more than 12 subjects respond, then stop the trial. Otherwise, proceed to the third stage by recruiting an additional 37 subjects. We would conclude that the test drug is effective if there are more than 25 responses for the subjects in the three stage combined.

6.6.2 Multiple-Arm Trials

In the previous section, we introduced procedures for sample size calculation under (flexible) optimal multiple-stage designs for phase II cancer trials with single arm. Sargent and Goldberg (2001) proposed a flexible optimal design considering a phase II trial that allows

clinical scientists to select the treatment to proceed for further testing for a phase III trial based on other factors when the difference in the observed responses rates between two treatments falls into the interval $[-\delta, \delta]$, where δ is a prespecified quantity. The proposed rule is that if the observed difference in the response rates of the treatments is larger than δ , then the treatment with the highest observed response rate is selected. On the other hand, if the observed difference is less than or equal to δ , other factors may be considered in the selection. In this framework, it is not essential that the very best treatment is definitely selected; rather, it is important that a substantially inferior treatment is not selected when a superior treatment exists.

To illustrate the concept proposed by Sargent and Goldberg (2001), for simplicity, consider a two-arm trial. Let p_1 and p_2 denote the true response rates for the poor treatment and the better treatment, respectively. Without loss of generality, assume that $p_2 > p_1$.

Let \hat{p}_1 and \hat{p}_2 denote the corresponding observed response rates for treatment 1 and treatment 2, respectively. Sargent and Goldberg (2001) considered the probability of correctly choosing the better treatment, i.e.,

$$P_{Corr} = P\{\hat{p}_1 > \hat{p}_2 + \delta | p_1, p_2\}$$

and the probability of the difference between the two observed response rates falling into the ambiguous range $[-\delta, \delta]$, i.e.,

$$P_{Amb} = P\{-\delta \leq \hat{p}_2 - \hat{p}_1 \leq \delta | p_1, p_2\}$$

Assuming that each treatment arm has the same number of subjects (i.e., $n_1 = n_2 = n$). The above two probabilities are given by

$$P_{Corr} = \sum_{x=0}^n \sum_{y=0}^n I\{(x-y)/n > \delta\} \binom{n}{x} \binom{n}{y} p_2^x (1-p_2)^{n-x} p_1^y (1-p_1)^{n-y}$$

and

$$P_{Amb} = \sum_{x=0}^n \sum_{y=0}^n I\{-\delta \leq (x-y)/n \leq \delta\} \binom{n}{x} \binom{n}{y} p_2^x (1-p_2)^{n-x} p_1^y (1-p_1)^{n-y}$$

where $I\{\text{inequality}\}$ is the indicator function, which equals 1 if the condition of the inequality holds and equals 0 otherwise. Sargent and Goldberg (2001) suggest n be selected such that $P_{corr} + \rho P_{Amb} > \gamma$, a prespecified threshold. Table 6.6.1 provides results for $\rho = 0$ and $\rho = 0.5$ for different sample sizes for $p_2 = 0.35$ and $\delta = 0.05$.

Table 6.6.1 Probability of Various Outcomes for Different Sample Sizes ($\delta = 0.05$)

n	p_1	p_2	P_{Corr}	P_{Amb}	$P_{Corr} + 0.5P_{Amb}$
50	0.25	0.35	0.71	0.24	0.83
50	0.20	0.35	0.87	0.12	0.93
75	0.25	0.35	0.76	0.21	0.87
75	0.20	0.35	0.92	0.07	0.96
100	0.25	0.35	0.76	0.23	0.87
100	0.20	0.35	0.94	0.06	0.97

Liu (2002) indicated that by the Central Limit Theorem, we have

$$p_{Corr} \approx P\left\{Z > \frac{\delta - \varepsilon}{\sigma}\right\}$$

and

$$p_{Amb} \approx P\left\{Z \leq \frac{\delta - \varepsilon}{\sigma}\right\} - P\left\{Z \leq \frac{-\delta - \varepsilon}{\sigma}\right\}$$

where Z is the standard normal variable, $\varepsilon = p_2 - p_1$ and

$$\sigma^2 = \frac{1}{n}[p_1(1-p_1) + p_2(1-p_2)]$$

The power of the test for the following hypotheses:

$$H_0: p_1 = p_2 \quad \text{vs.} \quad H_a: p_1 \neq p_2$$

is given by

$$1 - \beta = 1 - \Phi\left(Z(\alpha/2) - \frac{\varepsilon}{\delta}\right) + \Phi\left(-Z(\alpha/2) - \frac{\varepsilon}{\delta}\right),$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal variable and $Z(\alpha/2)$ is the $\alpha/2$ th upper quantile of the standard normal distribution.

Let $\lambda = p_{Corr} + p_{Amb}$. It can be verified that

$$\lambda = 1 - \Phi\left(Z(\alpha/2) - \frac{\varepsilon}{\delta}\right) + \rho\beta$$

As a result, sample size per arm required for a given λ can be obtained. Table 6.6.2 gives sample sizes per arm for $\delta = 0.05$ and $\lambda = 0.80$ or 0.90 assuming $\rho = 0$ or $\rho = 0.5$ based on exact binomial probabilities.

Note that the method proposed by Sargent and Goldberg (2001) can be extended to the case in which there are three or more treatments. The selection of the best treatment, however, may be based on pairwise comparison or a global test. Table 6.6.3 provides sample sizes per arm with three or four arms, assuming $\delta = 0.05$ and $\lambda = 0.80$ or 0.90 .

Table 6.6.2 Sample Sizes per Arm for Various λ Assuming $\delta = 0.05$ and $\rho = 0$ or $\rho = 0.5$

p_1	p_2	$\rho = 0$		$\rho = 0.5$
		$\lambda = 0.90$	$\lambda = 0.80$	
0.05	0.20	32	13	16
0.10	0.25	38	15	27
0.15	0.30	53	17	31
0.20	0.35	57	19	34
0.25	0.40	71	31	36
0.30	0.45	73	32	38
0.35	0.50	75	32	46
0.40	0.55	76	33	47

Table 6.6.3 Sample Size per Arm for Trials with Three or Four Arms for $\epsilon = 0.15$, $\delta = 0.05$, and $\lambda = 0.80$ or 0.90

ϵ	$n(\rho = 0)$		$n(\rho = 0.5)$	
	$r = 3$	$r = 4$	$r = 3$	$r = 4$
$\lambda = 0.80$				
0.2	18	31	13	16
0.3	38	54	26	32
0.4	54	73	31	39
0.5	58	78	34	50
$\lambda = 0.90$				
0.2	39	53	30	34
0.3	77	95	51	59
0.4	98	119	68	78
0.5	115	147	73	93

Chen and Ng (1998) indicated that the optimal two-stage design described above is similar to the Pocock sequence design for randomized, controlled clinical trials where the probability of early termination is high, the total possible sample size is larger, and the expected size under the alternative hypotheses is smaller (see also Pocock, 1997). The minimax design, on the other hand, is similar to O'Brien–Fleming design where the probability of early termination is low, the total possible size is smaller, but the expected size under the alternative hypotheses is larger (O'Brien and Fleming, 1979). The minimax design is useful when the patient source is limited, such as a rare cancer or a single site study.

Recently, multiple-stage designs have been proposed to monitor response and toxicity variables simultaneously. See, for example, Conaway and Petroni (1996), and Thall et al. (1995, 1996). In these designs, the multivariate outcomes are modeled and family-wise errors are controlled. It is suggested that this form of design should be frequently used in cancer clinical trials because delayed toxicity could be a problem in phase II trials. Chen (1997) also pointed out that one can use optimal three-stage design for toxicity monitoring (not simultaneous with the response). The role of *response* with that of *no toxicity* can be exchanged, and the designs are similarly optimal and minimax.

In practice, the actual size at each stage of the multiple-stage design may deviate slightly from the exact design. Green and Dalberg (1992) reviewed various phase II designs and modified each design to have variable sample sizes at each stage. They compared various flexible designs and concluded that flexible designs work well across a variety of p_0 and p_1 s, and powers.

6.7 RANDOMIZED PHASE II DESIGNS

In addition to optimal and flexible multiple-stage design introduced in the previous section, other designs such as Gehan's phase II design (Gehan, 1961) or Fleming's multiple-stage design (Fleming, 1982) had also been proposed to investigate anti-tumor activity during phase II clinical development of new promising agents. Mariani and Marubini (1996) provided a review of designs for phase II cancer clinical trials. However, most of these designs are single-arm, nonrandomized in nature, and without a comparative or

control group. In addition, the primary endpoint for searching promising anti-tumor agents during phase II development is the tumor response rate, possibly by different criteria. However, Simon et al. (1985) indicated that a tremendous variation exists in the observed response rates among different trials of the same anti-cancer agent. They also provided the following factors that contribute to the variability of observed response rates:

1. Patient selection
2. Response criteria
3. Interobserved variability in response assessment
4. Dosage modification and protocol compliance
5. Reporting procedures
6. Sample sizes

As a matter of fact, the response rate observed in a single-arm, nonrandomized cancer trial not only is a measure of anti-tumor activity of the new agent under investigation, but also it is a complicated function of the above-mentioned factors and their interactions with the new agent. Hence, it is extremely difficult or impossible to obtain an unbiased estimate of anti-tumor activity of the new agent based on tumor response rate from single-treatment, nonrandomized trials. Therefore, if several new treatments are available for evaluation of their anti-tumor activity, Simon et al. (1985) proposed the randomized phase II cancer clinical trials to select a promising new treatment for further evaluation in phase III clinical development. New treatments may consist of different new cytotoxic agents or involve analogs, different dosing regimen of the same agent, different preparations of the same drug, or different agonists or antagonists of a pharmaceutical compound.

The objective of phase II cancer clinical trials, usually based on the tumor response rate, is to select a promising new cytotoxic agent, among several available, with sufficient anti-tumor activity for further investigation. It follows that detection of a statistically significant difference among treatments by traditional hypothesis testing procedures not only fails to achieve the goal of phase II cancer clinical trials, but also it requires the same sample sizes as that of phase III trials. On the other hand, the statistical methods for ranking and selection can be directly applied to choose a promising new agent for phase III evaluations. In addition, ranking and selection do not directly test whether differences between treatments are statistically significant. Therefore, as will be shown below, the required sample size is much smaller than that required for phase III trials, and hence, it is reasonably manageable for any phase II cancer trials. Because for randomized phase II cancer clinical trials patients are randomly assigned to treatments, known or unknown factors that influence evaluation of tumor response rates are distributed approximately evenly or balanced among treatment groups. As a result, unbiased estimation of differences on the degree of tumor activity among treatment groups can be obtained, although due to the small sample size, the precision of such estimates may be not as high as that obtained from phase III trials. However, unbiasedness for comparison of tumor activity ensures that new agents can be reliably ranked when large differences are obtained.

The procedure for randomized phase II cancer clinical trials proposed by Simon et al. (1985) is rather simple. Initially, patients who meet inclusion and exclusion criteria are randomly assigned to one of treatments. When the study is completed, the promising new agent for further phase III evaluation is selected as the one with the greatest tumor response rate (or the best value of the other primary endpoint), regardless of how small or

nonsignificant its advantage over the other agents seems to be. The sample size for randomized phase II cancer clinical trial is determined to ensure that if a new agent is superior to all other test agents by an amount of δ , then it will be selected with a probability P , usually 0.8 or 0.9. If there are two new agents with equal anti-tumor activity that is superior to all others by δ , then the sample size should be determined to ensure that one of the two better agents will be chosen with a probability P . Reasonable values for δ are 0.15 or 0.20. $\delta < 0.15$ usually leads to an unrealistically large sample size, whereas $\delta > 0.20$ provides little or no information of the new agents (Simon and Hall, 1997). Table 6.7.1 provides the number of patients per arm for $\delta = 0.15$ and $P = 0.9$. Sample sizes for other conditions can be easily calculated according to the formulas given in the Appendix of Simon et al. (1985).

Example 6.7.1 Suppose that a sponsor intends to select one of the two new cytotoxic agents for further phase II evaluation. Based on historical information, it is expected that the smallest response rate of the two agents is about 30%. From Table 6.7.1, the sample size of 35 patients per group can provide a 90% probability of selecting the promising new agent that has a true response rate of 45%. However, the sample size for detection of a clinically meaningful difference of 15% (30% vs. 45%) with 90% power at the 5% significance level for a two-sided test is 230 per arm (Table A.3, Fleiss, 1981), a 6.6-fold increase from that required for ranking and selection of randomized phase II cancer clinical trials. However, as discussed before, a new agent will be selected and even the actual response rate of the better agent is less than 30% or the observed difference is smaller than 15%.

One of the reasons causing variation of tumor response rates is the inherent heterogeneity of cancer. For example, it is extremely difficult to evaluate tumor response of patients with prostatic, brain, or pancreatic cancers. On the other hand, for advanced colorectal cancer, advanced non-small-cell lung cancer, or advanced hepatocellular carcinoma, survival can be observed in a relatively short period of time. In such situations, survival rather than tumor response rate may be used as the primary endpoint for ranking and selection in the randomized phase II cancer clinical trials. As a new agent may not exist under which survival function is uniformly higher than those of all others across time, it is not clear how to define the best treatment based on observed survival functions. Liu et al. (1993), under the assumption of the Cox's proportional hazards model, suggested the use of hazard ratios for

Table 6.7.1 Number of Patients per Arm for Randomized Phase II Cancer Clinical Trials for $\delta = 0.15$ and $P = 0.9$

Smallest Response Rate(%)	Number of Treatments		
	2	3	4
10	21	31	37
20	29	44	52
30	35	52	62
40	37	55	67
50	36	54	65
60	32	49	59
70	26	39	47
80	16	24	29

Source: Simon et al. (1985).

ranking and selection of the promising new agents for further phase III trials. They also provided the required sample size for the least favorable configuration in which only one new agent has the best survival and the rest of the agents are all second best with the same survival. Table 6.7.2 presents the total sample sizes with $P = 90\%$ under the exponential distribution.

As the randomized phase II design has gained popularity in cancer research, its misuse and misapplications also become more frequently prevalently. Liu et al. (1999) indicated that the major misuse of the design is the treatment of the phase II trials as ends in themselves without further and definitive evaluation in more rigorous phase III trials. In addition, many practitioners also inappropriately perform post-hoc hypothesis testing and present p-values that are less than 0.05. Many users should remember that although the randomized phase II cancer clinical trials are randomized studies, the objective is to select a new promising agent with the best observed response rate (or survival) through ranking the observed response rates among all test agents. The sample size required for the randomized phase II cancer clinical trials will not be large enough to provide a definitive inference of superiority, noninferiority, or equivalence among different agents. Liu et al. (1999) reported that false-positive rates of the randomized phase II designs range from 20% to 40%. They especially cautioned against inclusion of control arms in the randomized phase II designs. If a control arm is included in the randomized phase II design of $K - 1$ new cytotoxic agents and if there is no difference among all treatments, then all treatments have equal chances of being the best agents, including the control. The probability that a new agent will appear better than the control is $(K - 1)/K$. When $K = 4$, this probability is 0.75. The probability of observing an impressive difference between the test agents and control is at least $2/K$ of the above probabilities. In this situation, if the control is selected because of its higher observed response over all other test agents, its effect can be disastrous. Liu et al. (1999) pointed out that a new test agent with similar efficacy as the control but less severe toxicities can be discarded as ineffective once and for all. Therefore,

Table 6.7.2 Total Sample Size for Randomized Phase II Cancer Clinical Trials Based on Survival for $P = 0.9$

Median	K	C = Ratio of Median Survivals								
		Two Groups			Three Groups			Four Groups		
		1.3	1.4	1.5	1.3	1.4	1.5	1.3	1.4	1.5
0.5	0	229	143	102	513	320	226	822	512	361
	0.5	141	88	62	317	197	138	598	315	221
	1	117	72	51	263	162	114	421	260	182
0.75	0	306	192	137	685	429	304	1098	686	485
	0.5	177	111	79	398	248	175	638	397	280
	1	139	87	61	312	194	136	500	310	218
1	0	384	242	173	860	539	383	1377	862	611
	0.5	216	135	96	483	302	214	774	483	341
	1	163	102	72	365	228	161	586	365	257

Source: Liu et al. (1993).

Median: median survival for all other groups.

K: Additional follow-up in years after the end of accrual.

C: Ratio of median survivals of the best agent to that of all other groups.

a control arm should not be included in the cancer clinical trials using the randomized phase II design.

6.8 DISCUSSION

A vast amount of literature in designs for phase I cancer trials is available. Our review in this chapter by no means is either comprehensive or complete. We, however, covered four fundamental and commonly used designs for phase I cancer clinical trials. These include the standard dose escalation design, one-stage or two-stage up-and-down designs, accelerated titration designs, and continual reassessment method. Table 6.8.1 summarized the basic characteristics of these designs. From Table 6.8.1, although the standard dose escalation design and most up-and-down designs are implicitly based on a dose-response model, they all select the MTD operationally without actually modeling the relationship between the occurrence of DLT and dose based on the data collected in the trial. For these designs, the traditional likelihood approaches such as delta method, the Fieller's theorem, and likelihood ratio method fail to provide an adequate estimate of the MTD and its corresponding confidence interval. The reason is that they are unable to take into account the adaptive and sequential nature of a discrete Markov chain for these designs (Storer, 1993). Storer (1989, 1993, 2001) proposed a two-parameter logistic regression mode to fit the data collected from design BD. On the other hand, as mentioned above, the CRM employs a single-parameter model to fit the data after the completion of the study. It should be noted that for the CRM, the model for updating the dose levels is different from the model for the dose-response model to describe the relationship between the DLT and the dose. In addition, Shen and O'Quigley (1996) showed that the CRM provides a consistent estimate of the MTD even under model misspecification.

Over the last decade, many simulations have been performed to empirically compare the standard dose escalation design, up-and-down designs, the original CRM, and its various modifications. The results can be found in O'Quigley and Chevret (1991), Korn et al. (1994), Goodman et al. (1995), Ahn (1998), O'Quigley (1999), Korn et al. (1999), and Storer (2001). Some of the results are summarized below:

1. The standard dose escalation design treats more patients at the subtherapeutic dose levels.
2. The standard dose escalation design underestimates the maximum tolerable dose.
3. The original CRM requires fewer patients than does the standard dose escalation design.
4. The average number of cohorts of the original CRM with a patient per cohort is larger than that of the standard dose escalation design. Hence, the duration of the trials using the original CRM may be longer than other phase I designs.
5. The average number of cohorts reduces dramatically for the modified CRM with three patients per cohort and is similar to that of the standard dose escalation design.
6. The two-stage (modified) CRM does not provide better performance than the one-stage modified CRM.
7. The CRM is independent of the targeted percentile of some tolerance distribution that is prespecified for other designs. In addition, it, theoretically, has convergence properties.

Table 6.8.1 Summary of Characteristics for Phase I Designs

Characteristics	Standard Dose Escalation Design	Up-and-Down Designs	ATD	CRM
Initial dose for the first patient	Lowest dose	Not necessarily lowest dose	Lowest dose	Not necessarily lowest dose
Dosing interval	Single step	Single step	Single step	Multiple steps
Deescalation	No	Yes	No	Yes
Intra-patient dose modification	No	No	Yes	No
Differential toxicity	No	No	Yes	No
DLT by the first course	Yes	Yes	Not necessary	Yes
# of patients per cohort	3	1–3	1–3	1–3
Fixed sample size	No	Possible	No	Yes
MTD estimated from pre-specified dose level	Yes	Not necessary	Yes	Yes
Modeling fitting for estimation of MTD	No	Possible	Possible	Yes

8. If the number of patients is not fixed in advance, design BD required more cohorts than the standard dose escalation design.
9. Design BD exposes more patients at toxic doses than the standard dose escalation design.
10. Design BD provides improved estimates of MTD over the standard dose escalation design.
11. No design performs uniformly well in all possible dose-response settings.
12. The estimates of MTD generated from the CRM generally have smaller bias than those from design BD, although the bias is relatively small.
13. The estimates of MTD from design BD have a greater precision than those of the CRM. The addition of a model-fitting step only slightly improves the precision of CRM.
14. The standard dose escalation design produces the estimates of MTD with largest bias and worst precision.

However, little or no empirical evidence is available on evaluation of performance between the accelerated titration design and the CRM. This urgently deserves further investigation.

Although the primary objective of phase II cancer clinical trials is to select a new anti-cancer agent for which its anti-tumor activity reaches a prespecified target level or above, it is not uncommon that phase II cancer clinical trials try to evaluate a new anti-cancer treatment against the current standard treatment. This type of trials is referred to as phase IIB cancer clinical trials (Simon and Hall, 1997). Phase II cancer clinical trials are also conducted without inclusion of the standard therapy in a single-treatment and nonrandomized fashion despite the fact its objective is to compare the anti-tumor activity between the new and standard treatments. One way to encounter this issue is to let p_0 in hypothesis (6.1.1) represent the anti-tumor activity of the standard treatment. However, p_0 is a population parameter that must be estimated from the data provided by the trials conducted previously for evaluation of the standard treatment. Any inference of the new treatment by assuming the estimate of p_0 as a constant is incorrect because the variability of estimate of p_0 is not accounted for. Thall and Simon (1990) developed optimal single-stage phase II designs that use historical data from the previous trials of the standard treatment to account for the variability of p_0 of prior studies. On the other hand, Fazzari et al. (2000) proposed another method to taking into account the variability inherent in estimation by setting p_0 as the $(1 - \alpha)100\%$ upper confidence limit for the k-year survival rate derived from the historical data. Thall and Simon (1990) also found that a trial with an imbalanced randomization to both new and standard treatments might be superior to a single-arm, nonrandomized trial of the new treatment alone. In addition, they concluded that when the uncertainty of estimates for anti-tumor activity in historical control is great, the traditional phase IIB cancer clinical trial is not reliable. The Bayesian approach to phase II cancer clinical trials has been suggested. For details, see Thall and Simon (1994a, 1994b).

7

CLASSIFICATION OF CLINICAL TRIALS

7.1 INTRODUCTION

For approval of a new drug, the FDA requires that substantial evidence of the effectiveness of the drug be provided through the conduct of adequate and well-controlled clinical trials. The characteristics of an adequate and well-controlled clinical study include appropriate methods for bias reduction such as double-blinding, randomization of treatment assignments, a well-defined patient population, and scientific and valid statistical methods for data analysis.

Basically, there are several different types of clinical trials, depending on their functions and performance characteristics, although they are not mutually exclusive. These different types of clinical trials include multicenter trials, superiority trials, sequential trials, active control and equivalence/noninferiority trials, dose-response trials, combination trials, bridging studies, and vaccine clinical trials. They are usually applied in different situations depending on the objectives of the planned clinical trials. For example, a *multicenter trial* may be desirable because it provides replication and generalizability of clinical results to the target patient population across study centers. If the objective of the intended clinical trial is to show that the test treatment is better than a concurrent control in terms of its primary clinical endpoints, then it is referred to as the *superiority trial*. In many cases, an *active control trial* is necessarily conducted to establish the efficacy of a new drug when the patients under study are very ill or have severe or life-threatening diseases. On the other hand, one of the goals for the active control trials may be to verify that the efficacy of the test treatment is similar or is not worse than that of the current active control. The active control trials with the objective of this kind are referred to as *equivalence/noninferiority*

trials. Noninferiority trials are to establish therapeutic equivalence between the test treatment and an active control by showing its noninferior efficacy with respect to the active control (usually the current standard treatment). For the approval of a generic drug, a *bioequivalence trial* is required by the FDA to demonstrate that the generic copy is bioequivalent to the innovator drug product in terms of the rate and extent of drug absorption. Bioequivalence is usually established by showing that the pharmacokinetic parameters derived from the blood or plasma concentration time curve of the active ingredient of the generic copy is similar to those from the innovator drug product.

During the phase II clinical development, *dose-response trials* are usually conducted to determine the therapeutic window of the test drug and the relationship between the efficacy/safety and doses within the therapeutic window. When the drug under study consists of more than one active ingredient, a *combination trial* is required to assess the treatment effect by taking into account potential drug-to-drug interaction. Combination trials are particularly useful in evaluation of AIDS and cancer treatments or alternative herbal medicines. To address the impact of ethnic factors, the ICH E5 Guideline entitled, *Ethnic Factors in the Acceptability of Foreign Clinical Data*, suggests conducting *bridging studies* in the new region to extrapolate the foreign data to the new region. Because the target patient population for *vaccine trials* is usually normal healthy individuals, ethnics, safety, design, and sample size for vaccine trials require special considerations.

In the next section, the limitations of single-site studies and the feasibility of multicenter trials are briefly discussed. The concept and considerations for superiority trials are given in Section 7.3. Issues and concerns for the use of active control and equivalence/noninferiority trials are provided in Section 7.4. Section 7.5 presents the basic features and characteristics of dose-response trials. Some statistical deliberations for combination trials are outlined in Section 7.6. Objectives of bridging studies and their dilemmas and challenges are given in Section 7.7. Basic design considerations and statistical methods for vaccine trials are discussed in Section 7.8. Final remarks and discussion are presented in Section 7.9.

7.2 MULTICENTER TRIAL

When conducting a clinical trial, it may be desirable to have the study done at a single study site if (1) the study site can provide an adequate number of relatively homogeneous patients that represent the targeted patient population under study and (2) the study site has sufficient capacity, resources, and supporting staff to sponsor the study. One of the advantages for a single-site study is that it provides consistent assessment for efficacy and safety in a similar medical environment. As a result, a single-study site can improve the quality and reliability of the collected clinical data and consequently the inference of the clinical results. However, a single-site study has its own limitations and hence may not be feasible in many clinical trials. These limitations include the availability of patients and resources in a single site. If the intended clinical trial calls for a large number of patients, a single-site study may take a long time to complete the study, since qualified patients may not be available at the same time. Besides, even if qualified patients were available at the same time, the single-study site may not have sufficient resources to enroll these patients at the same time. In practice, qualified patients are usually enrolled sequentially at different times until the required number of patients is reached to achieve a desired power for the detection of a clinically meaningful difference.

Goldberg and Kury (1990) indicate that a single-site study may not be appropriate in situations where (1) the intended clinical trials are for relatively rare chronic diseases and (2) the clinical endpoints for the intended clinical trials are relatively rare (i.e., require a large number of patients to observe an incidence). For example, as observed by Goldberg and Kury (1990), if the intended clinical trial is to study a relatively rare chronic disease such as polycythemia vera (a disease characterized by an elevated hematocrit which has as natural consequences, stroke, hemorrhage, leukemia, and death), a single-site study is not feasible because it is unlikely that a single site is able to recruit a sufficient number of patients within a relatively short time frame to achieve the desired power for the detection of a meaningful clinical difference. Even if the single site is able to recruit the required number of patients, it will take a relatively longer time to complete the study. As Goldberg and Kury (1990) point out, if the clinical endpoint for the intended clinical trial is relatively rare such as mortality in clinical trials for acute myocardial infarction, then the single site may be required to enroll a large number of patients in order to observe a mortality. In such case, a single-site study is of little practical interest.

To overcome the disadvantages of a single-site study, the multicenter study is usually considered. A multicenter study is a single study involving several study centers (sites or investigators). In other words, a multicenter trial is a trial conducted at more than one distinct center where the data collected from these centers are intended to be analyzed as a whole. At each center an identical study protocol is used. A multicenter trial is a trial with a center or site as a natural blocking or stratified variable that provides replications of clinical results. A multicenter trial should permit an overall estimation of the treatment difference for the targeted patient population across various centers. In what follows, we will discuss the impact of treatment-by-center interaction and some practical issues when planning a multicenter trial.

Treatment-by-Center Interaction

The FDA guideline suggests that individual center results should be presented for a multicenter study. In addition, the FDA suggests that statistical tests for homogeneity across centers (i.e., for detecting treatment-by-center interaction) be provided. The significant level used to declare the significance of a given test for a treatment-by-center interaction should be considered in light of the sample sizes involved. Any extreme or opposite results among centers should be noted and discussed. For the presentation of the data, demographic, baseline, and postbaseline data as well as efficacy data should be presented by center, even though the combined analysis may be the primary one. Gail and Simon (1985) classify the nature of interaction as either quantitative or qualitative. A quantitative interaction between treatment and center indicates that the treatment differences are in the same direction across centers but the magnitude differs from center to center, while a qualitative interaction reveals that substantial treatment differences occur in different directions in different centers. Figure 7.2.1 depicts situations where there are quantitative and qualitative treatment-by-center interactions. As an illustration, consider the following two examples which exhibit quantitative and qualitative treatment-by-center interactions.

As indicated by Ebbeling and Clarkson (1989), muscle soreness and elevations in serum creatine kinase (CK) can be used as a biochemical marker for skeletal muscle injury. Recently a study was conducted to evaluate the hypothesis that exposure of skeletal muscle to exercise-induced stress in combination with prior administration of a study drug (e.g., drug A) will produce a greater increase in CK than the effects of exercise alone. This

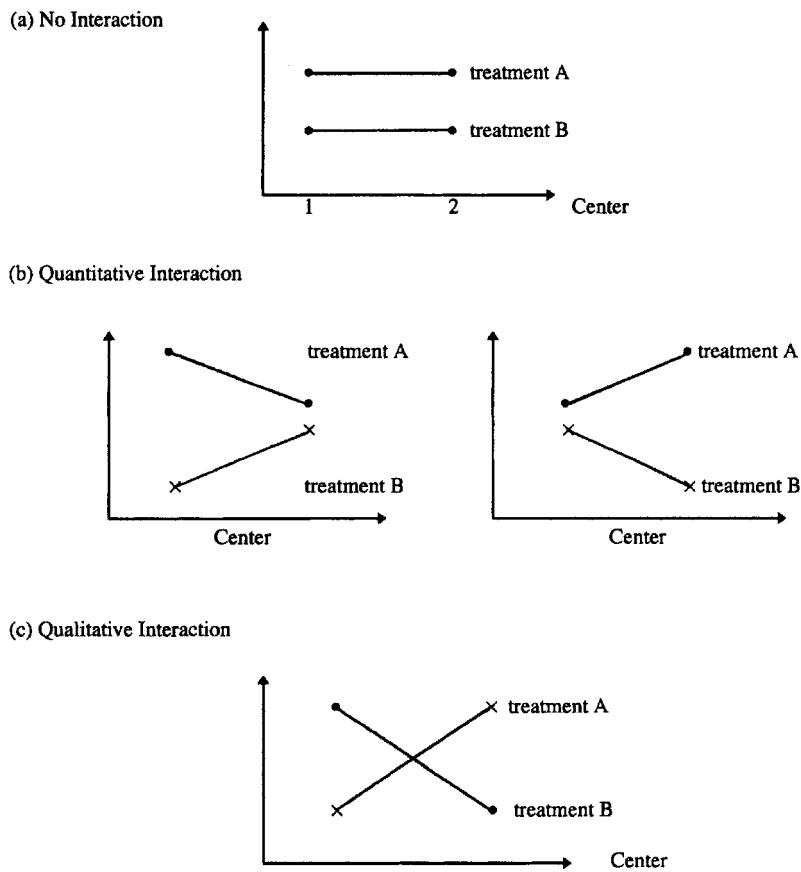


Figure 7.2.1 Treatment-by-center interaction

study was a randomized, two-arm parallel study comparing the study drug A plus exercise and a placebo control plus exercise which was conducted at two distinct study centers. Table 7.2.1 lists the mean CK values at 24 hours post-treatment. Figure 7.2.2 provides a preliminary investigation of a potential treatment-by-center interaction. The plot shows that there is a potential quantitative treatment-by-center interaction. This was confirmed by an analysis of variance that includes the terms of treatment, the center, and the treatment by center interaction (p -value = 0.08 < 0.1). A subgroup analysis by center reveals that there is a significant treatment effect at the second center (p -value = 0.029 < 0.05).

As another example, consider a clinical trial for the assessment of a study drug's efficacy in treating mild to moderate hypertension. The study was conducted as a randomized placebo-controlled multicenter trial that involved 219 patients in 27 study centers. Table 7.2.2 lists the mean change in seated diastolic blood pressures after six weeks of treatment. A plot of mean change in seated diastolic pressures against study centers is given in Figure 7.2.3. In the figure the centers are grouped according to the magnitude of differences in mean change from the baseline. That is, the center with a difference in mean change from baseline of -19.89 are labeled site 1 and the center with a difference in mean change from baseline of 9.11 are labeled site 27. As can be seen, 19 centers (70.4%) show the difference

Table 7.2.1 Mean CK Values at 24 Hours Post-Treatment

Treatment	Center	
	1	2
Drug A	15 355.87 (69.85)	9 549.78 (237.27)
Placebo	16 213.31 (26.34)	11 208.64 (35.79)

Note: The top values and the values in the parentheses are corresponding sample sizes and standard errors.

in the positive direction, while 8 centers (29.6%) are in the negative direction. An analysis of variance indicates that a significant qualitative interaction between treatment and group has occurred (p -value = 0.01). If we ignore the interaction and perform an analysis of variance, an overall estimate of the treatment effect based on the difference in the mean change from the baseline is given by -4.36 ± 1.9 or $(-6.26, -2.46)$ with a p -value less than 0.01. This positive significant result, however, is somewhat misleading because it is not reproducible in those 8 out of 27 centers. In other words, there is a relatively high chance that we may observe a totally opposite result if we are to randomly select a center from a pool of centers and repeat the study with the same protocol. Besides, the 8 centers involve 60 patients, who show different results, or about 27.4% of the patients under study. Therefore the reproducibility and generalizability of the results to the targeted patient population and the treatment setting is questionable.

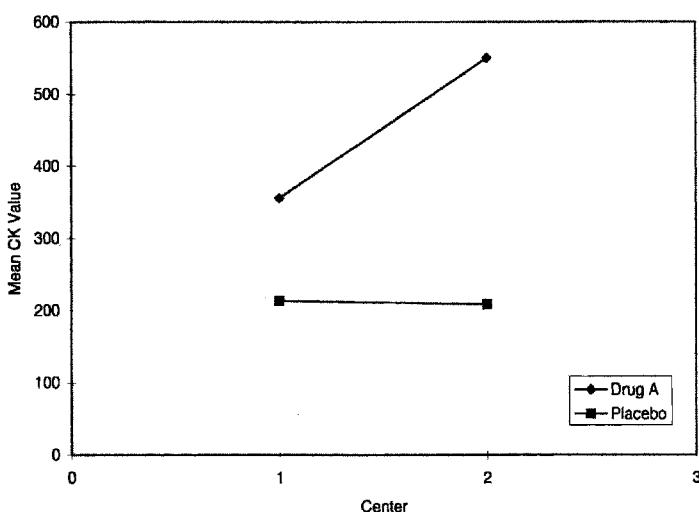


Figure 7.2.2 The mean CK values at 24 hours post-treatment

Table 7.2.2 Mean Change From the Baseline for Seated Diastolic Blood Pressure After Six Weeks of Treatment

Site	<i>N</i>	Placebo	Drug	Difference
1	6	2.44	-17.44	-19.88
2	9	1.07	-13.00	-14.07
3	29	-1.74	-12.07	-10.33
4	5	-2.44	-11.73	-9.29
5	5	-1.70	-9.24	-7.54
6	5	-0.67	-7.89	-7.22
7	12	-3.89	-11.06	-7.17
8	4	-11.00	-18.00	-7.00
9	6	1.33	-4.67	-6.00
10	24	-1.18	-6.77	-5.59
11	7	-8.00	-13.50	-5.50
12	7	-6.44	-11.75	-5.31
13	8	-4.83	-10.00	-5.17
14	4	-1.00	-5.93	-4.93
15	6	-6.44	-11.33	-4.89
16	7	-7.83	-12.44	-4.61
17	8	-6.17	-10.47	-4.30
18	4	-2.07	-5.67	-3.60
19	11	-4.56	-7.93	-3.37
20	6	-11.78	-11.67	0.11
21	5	-8.67	-7.53	1.14
22	11	-14.67	-13.20	1.47
23	6	-14.22	-12.00	2.22
24	8	-13.67	-10.50	3.17
25	9	-8.00	-2.80	5.20
26	10	-8.27	0.67	8.94
27	5	-13.11	-4.00	9.11
Total	219	-5.50	-9.78	-4.36

Practical Issues

As was indicated earlier, a multicenter trial with a number of centers is often conducted to expedite the patient recruitment process. Although these centers usually follow the same study protocol to evaluate the efficacy and safety of a study drug, some design issues need to be carefully considered. These design issues include the selection of centers, the randomization of treatments, and the use of a central laboratory for laboratory evaluations. The selection of centers is important to constitute a representative sample for the targeted patient population. However, in multicenter trials the centers are usually selected based on convenience and availability. When planning a multicenter trial with a fixed sample size, it is important to determine the allocation of the centers and the number of patients in each center. For example, if the intended clinical trial calls for 100 patients, the sponsor may choose to have 5 study centers with 20 patients in each, 10 study centers with 10 patients in each, or 20 study centers with 5 patients in each. The chance for observing a significant treatment-by-center interaction for the selection of 20 centers is expected to be higher than those for the selections of 10 centers and 5 centers. If there are potential dropouts, the

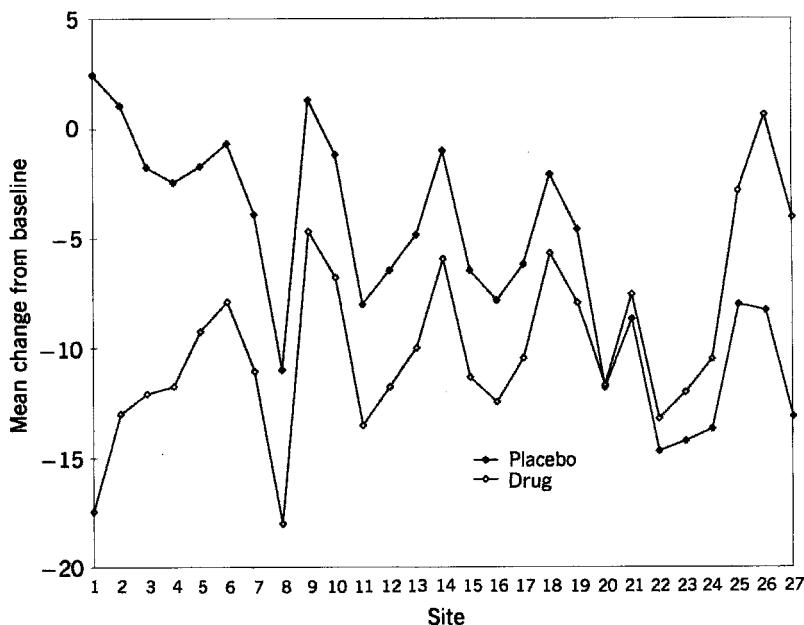


Figure 7.2.3 Mean seated diastolic blood pressures versus study site.

selection of 20 centers may result in a number of small centers (i.e., with a few patients in the center). For comparative clinical trials, the comparison between treatments is usually made between patients within centers. If there is no treatment-by-center interaction, the data can be pooled for analysis across centers. Therefore, it is not desirable to have too few patients in each center. A rule-of-thumb is that the number of patients in each center should not be less than the number of centers. In this case the selection of 10 sites for a fixed sample size of 100 patients may be preferable. Some statistical justification for this rule can be found in Shao and Chow (1993). Once the centers are selected, it is also important to assign treatments to patients in a random fashion within each center. The methods of randomization described in Chapter 4 should be applied. The issue whether a central laboratory will be used for the laboratory testing of samples collected from different centers has a significant impact on the assessment of efficacy and safety of the study drug. A central laboratory provides a consistent assessment for laboratory tests. If a central laboratory is not used, the assessment of laboratory tests may differ from center to center depending on the equipment, analyst, and laboratory normal ranges used at that center. In such a case possible confounding makes it difficult to combine the laboratory values obtained from the different centers for an unbiased assessment of the safety and efficacy of the study drug.

Another practical issue of great concern in a multicenter trial is statistical analysis of the collected data from each center. As was indicated earlier, if there is no evidence of treatment-by-center interaction, the data can be pooled for analysis across centers. The analysis with combined data provides an overall estimate of the treatment effect across centers. In practice, however, if there are a large number of centers, we may observe significant treatment-by-center interaction, either quantitative or qualitative. As indicated by Gail and Simon (1985), the existence of a quantitative interaction between treatment and center does not invalidate the analysis in pooling data across centers. An overall estimate

of the average treatment difference is statistically justifiable, and it provides a meaningful summary of the results across centers. On the other hand, if a qualitative interaction between treatment and center is observed, the overall or average summary statistic may be misleading and hence considered inadequate. In this case it is preferable to describe the nature of the interaction and to indicate which centers contribute toward the interaction. In practice, if there are too many centers, the trial may end up with big imbalances among centers, in that some centers may have a few patients and others a large number of patients. If there are too many small centers (with a few patients in each center), we may consider the following two approaches. The first approach is to combine these small centers to form a new *center* based on their geographical locations or some criteria prespecified in the protocol. The data can then be reanalyzed by treating the created *center* as a regular center. Another approach is to randomly assign the patients in these small centers to those larger centers and reanalyze the data. This approach is valid under the assumption that each patient in a small center has an equal chance of being treated at a large center.

As was indicated before, for approval of a new drug, two well-controlled clinical trials are conducted to provide substantial evidence of the effectiveness and safety of the new drug. Since a multicenter trial involves more than one distinct center, whether a single multicenter trial is equivalent to that of two separate trials has become an interesting question. The FDA indicates that an *a priori* division of a single multicenter trial into two studies is acceptable for establishing the reproducibility of drug efficacy to NDA approval. Nevius (1988) proposes a set of four conditions under which evidence from a single multicenter trial would provide sufficient statistical evidence of efficacy. These conditions are summarized below:

1. The combined analysis shows significant results.
2. There is consistency over centers in terms of direction of results.
3. There is consistency over centers in terms of producing nominally significant results in centers with sufficient power.
4. Multiple centers show evidence of efficacy after adjustment for multiple comparisons.

To address the consistency over centers, Chinchilli and Bortey (1991) propose the use of the noncentrality parameter of an *F* distribution as a means of testing for consistency the treatment effect across centers. As an alternative, Khatri and Patel (1992) and Tsai and Patel (1992) consider a multivariate approach assuming random center effects. Huster and Louv (1992) suggest the use of the minimax statistic as a method that can quantify the amount of evidence for reproducibility of treatment efficacy in a single multicenter trial. To describe the minimax statistic, we consider an example given by Huster and Louv (1992). Suppose that there is a four-center clinical trial. Denote the four centers by *a*, *b*, *c*, and *d*. Consider all possible divisions of these four centers into two mutually exclusive sets of centers. The minimax statistic can then be summarized as follows: First, for each of the seven divisions, find the *p*-value for the drug effect (adjusted for the center) for each study. Then find the maximum of these two *p*-values for each division. Second, find the minimum of these maximum *p*-values across all divisions. The rationale for choosing the maximum at the first step is that if a particular division is to show reproducibility of a drug effect, then both studies within that division should exhibit a significant drug effect. On the other hand, the rationale for choosing minimum at the second step is that the optimal division is one where the evidence against the null hypothesis from both studies is the greatest. As indicated by Huster and Louv (1992), the minimax statistic approach provides a reasonable

assessment for the amount of evidence for reproducibility of treatment efficacy in a single multicenter trial. In addition it gives an objective answer to the question whether a second confirmatory study is required.

Note that the analysis of a multicenter trial is different from that of a meta-analysis. The analysis of multicenter trials combines data observed from each study center; the data are generated based on the methods prospectively specified in the same study protocol with the same method of randomization and probably at the same time. In contrast, a meta-analysis combines data retrospectively observed from a number of independent clinical trials, which may be conducted under different study protocols with different randomization schemes at different times. In either case the treatment-by-center interaction for multicenter trials or treatment-by-study interaction for meta-analyses must be carefully evaluated before pooling the data for analysis.

7.3 SUPERIORITY TRIALS

According to the ICH E9 Guideline entitled, *Statistical Principles for Clinical Trials*, a superiority trial is defined as a trial with the primary objective of showing that the response to the investigational product is superior to a comparative agent. Scientifically, superiority trials can provide the most convincing evidence of superior efficacy of the test drug product to that of comparative controls. However, as discussed in Section 2.6, from a regulatory review/approval process point of view, the superiority of a test drug product is in fact established in a two-step procedure. Let μ_T and μ_C be the summary population parameters of some primary efficacy endpoint for the test drug product and comparative control, respectively. Here μ_T and μ_C can be the population means if the primary efficacy endpoint is a continuous variable, or the proportions if the primary efficacy endpoint is a binary variable or survival function if the primary efficacy endpoint is a censored variable. In addition, a larger value of the summary population parameter represents a superior efficacy. Then, the first step is to establish the superior efficacy of the test drug product by testing the following two-sided hypotheses at a prespecified significance level, say, α :

$$H_0: \mu_T = \mu_C \quad \text{versus} \quad H_a: \mu_T \neq \mu_C. \quad (7.3.1)$$

After the null hypothesis is rejected at the α level of significance and if the estimate of $\mu_T - \mu_C$ is positive, the test drug product is then claimed to be superior to the control agent in efficacy. The ICH E9 guideline indicates that this two-step procedure is consistent with the two-sided confidence intervals that are generally considered an appropriate method for estimating the possible size of the difference between the test drug product and the comparative control. In other words, the superiority of the test drug product is established if the lower limit of the $(1 - \alpha)\%$ two-sided confidence interval for $\mu_T - \mu_C$ is greater than 0. This two-step procedure is actually equivalent to testing the following one-sided hypotheses at the $\alpha/2$ significance level:

$$H_0: \mu_T \leq \mu_C \quad \text{vs.} \quad H_a: \mu_T > \mu_C. \quad (7.3.2)$$

It follows that this two-step approach establishes the superiority of the test drug product at the $\alpha/2$ significance level rather than at the usual α significance level. Hence, the issue of one-sided or two-sided approaches to inference on establishment of superior efficacy

remains controversial and generates a lot of discussions and debates. The rationale behind this controversial issue from a regulatory viewpoint is that no one would know for sure during the clinical development that there is a difference between the test drug product and its comparative control. Therefore, one should first provide evidence of the difference between the test drug product and comparative control at the usual α significance level. If there is no such evidence, then no claim of superior efficacy for the test drug product can be concluded. Once sufficient evidence is provided to support the existence of a difference between the test drug product and the comparative control, then superiority of the test drug product can be claimed only when the difference is in the positive direction. Otherwise, the efficacy of the test drug product is inferior to the comparative control. For more details on this topic, see Peace (1991), Koch (1991), Fisher (1991), and Dubey (1991).

Example 7.3.1 Chelation Therapy for Ischemic Heart Disease

As indicated in Example 3.4.1, the PATCH is a double-blind, randomized, placebo-controlled trial to evaluate whether the EDTA chelation therapy is superior to placebo on exercise ischemic threshold and quality of life in patients with stable ischemic heart disease. The primary efficacy endpoints for this study are change from baseline to 27-week follow-up in time to ischemia (1-mm ST depression), and mental and physical component summary of Health Status Survey Short Form-36 (Ware et al., 1994). The sample size of 40 patients per group for this study was chosen to provide a 90% power for detection of a difference of 60 seconds in mean change in exercise time from baseline to 27-week follow-up, assuming a standard deviation of 80 seconds in both groups. The results of these two primary endpoints are summarized in Table 7.3.1. The p -values for treatment comparisons given in Table 7.3.1 are for the two-sided hypotheses given in (7.3.1). The mean change in exercise time to ischemia at the 27-week follow-up for the chelation and placebo groups were 63 seconds (95% CI: 29 to 95, p -value < 0.001) and 54 seconds (95% CI: 23 to 84, p -value < 0.001), respectively. Although improvement of the exercise time to ischemia is highly statistically significant for both groups, the difference in mean change at 27-week follow-up between the chelation and placebo group is only 9 seconds with a 95% confidence interval from -36 to 53 seconds. As the p -value for the two-sided hypothesis in (7.3.1) is 0.69 > 0.05 and the lower limit of the 95% confidence interval is -36 seconds < 0, then based on the exercise time to ischemia, there is no evidence to support that EDTA chelation therapy provides superior efficacy to placebo. A similar conclusion can be reached based on the results on the mental and physical component summaries of SF-36. Because the standard deviations of these primary efficacy endpoints and other variables were comparable between the chelation and placebo groups, this trial is considered a well-conducted trial.

As mentioned in the ICH E9 guideline for regulatory settings, the approach of setting type I errors for one-sided tests at half the conventional type I error used in two-sided test is preferable for superiority. However, such a requirement is not necessary for other research environments in which superiority can be proved in a one-step procedure by performing a one-sided hypothesis (7.3.2) at the traditional α level of significance.

Example 7.3.2 The Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial

In the United States, the most common cancers are lung and colorectal cancers that accounted for 46% of cancer deaths in males and 34% of cancer death in females in 1989. Prostate

Table 7.3.1 Summary of Results of the PATCH

<i>Endpoint: Time to Ischemia in Seconds</i>				
Time Points	Statistics	Chelation	Placebo	Difference
Baseline 27 Weeks	N	39	39	
	Mean (SD)	589 (176)	572 (172)	
	Mean (SD)	652 (174)	626 (186)	
	Mean Change	63 (29, 95)	54 (23, 84)	9 (-36, 53)
	p-value	<0.001	<0.001	0.69
<i>Endpoint: SF-36 Mental Component Summary</i>				
Time Points	Statistics	Chelation	Placebo	Difference
Baseline 27 Weeks	N	39	39	
	Mean (SD)	52.6 (7.6)	48.3 (10.4)	
	Mean (SD)	54.6 (6.7)	50.5 (9.2)	
	Mean Change	2.1 (-0.4, 3.6)	2.1 (-0.4, 4.5)	0.01 (-3.4, 3.4)
	p-value	0.10	0.09	0.99
<i>Endpoint: SF-36 Physical Component Summary</i>				
Time Points	Statistics	Chelation	Placebo	Difference
Baseline 27 Weeks	N	39	39	
	Mean (SD)	42.9 (10.1)	39.9 (11.0)	
	Mean (SD)	45.1 (10.0)	44.9 (10.7)	
	Mean Change	2.2 (-0.5, 4.9)	5.0 (2.7, 7.3)	-2.8 (-6.3, 0.6)
	p-value	0.11	<0.001	0.11

Source: Knudtson et al., 2002.

cancer was the third leading cause of cancer mortality in males and accounts for 11% of cancer deaths. Ovarian cancer accounted for 5% of cancer deaths in females. In short, these types of cancers account for approximately half of all diagnosed cancers and half of all cancer deaths in the United States. Although these cancers in males and females are clinically significant and have a huge impact in public health and health policy, uncertainty regarding the value of screening for these cancers has resulted in a conflicting position in the medical community and confusion in populations at risk. Therefore, a randomized, controlled clinical trial is necessarily conducted to determine the efficacy of screening based on disease-specific mortality. The Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial is a large-scale, long-term randomized, control trial conducted by the United States National Cancer Institute (NCI) to address the above issues (PLCO Project Team, 2000). Specifically, the objectives of the PLCO Cancer Screening Trial are in screenees aged between 55 and 74 years at entry, whether:

In Females and Males

1. Screening with flexible sigmoidoscopy (60-cm sigmoidoscope) can reduce mortality of 25% from colorectal cancer as compared to the control group of usual medical care.

2. Screening with chest X-ray can reduce mortality of 20% from lung cancer as compared to the control group of usual medical care.

In males

3. Screening with digital rectal examination (DRE) plus serum prostate-specific antigen (PSA) can reduce mortality of 25% from prostate cancer as compared to the control group of usual medical care.

In females

4. Screening with CA125 blood test and transvaginal ultrasound (TVU) can reduce mortality of 35% from prostate cancer as compared to the control group of usual medical care.

As pointed out by the PLCO Project Team (2000), unlike the drug regulatory setting, the question for each of the four types of cancer is not whether screening reduces or increases mortality. In addition, determination of whether screening increases mortality is simply not an objective of the trial. Furthermore, the PLCO Project Team (2000) also indicated that if the screening intervention has no effect or if it is harmful, the consequences in terms of a public health decision are the same—screening is not recommended. Therefore, whether screening reduces mortality as compared to the usual medical care is of research interest to the PLCO Project Team. Clearly, this is a superiority trial corresponding to the one-sided hypotheses described in (7.3.2), which involves a one-sided design and analysis approach. As a result, sample-size determination based on a one-sided hypotheses testing approach was employed. To provide approximately 90% power for detection of the above-mentioned magnitude of reduction in mortality for each of the four types of cancers, a total of 74,000 females and 74,000 males were enrolled. Half of the 74,000 females (37,000) were randomized to the screening group of chest X-ray, flexible sigmoidoscopy, CA125, and TVU, and the other half of them were assigned to the control group of the usual medical care. Half of the 74,000 males (37,000) were randomized to the screening group of chest X-ray, flexible sigmoidoscopy, DRE, and PSA, and the other half of them were assigned to the control group of the usual medical care. Each patient will be followed up for at least 13 years. The pilot phase of the study began in September 1992 for 2 years, and the main phase of study started in September 1994. The main phase of the study consists of an 8-year accrual period and a 5-year screening period and a 13-year follow-up period of each patient for a total duration of 23 years. The PLCO Cancer Screening Trial is scheduled to conclude by September 30, 2015.

7.4 ACTIVE CONTROL AND EQUIVALENCE/ NONINFERIORITY TRIALS

For approval of a test drug, it is required by regulatory agencies that clinical trials be conducted comparing the test drug with a control. Section 314.126 in Part 21 of CFR and ICH E10 Guideline entitled, *Choice of Control Group in Clinical Trials*, indicate that there are five kinds of control, including placebo concurrent control, dose-response concurrent control, active (positive) concurrent control, no treatment concurrent control, and external or historical control (ICH, 1999). An overview of these controls and examples are given in Section 3.4. A control is usually referred to as active or positive if it is a known active treatment or

drug. Therefore, an active control (positive control) trial is defined as an adequate and well-controlled trial in which a test drug product is compared concurrently with a known active drug. Here, an active control trial is referred to as a two-arm study in which subjects are randomly assigned to the test drug product or to the known active drug.

From a regulatory prospective, the effectiveness of a drug product should be established by demonstrating its superior efficacy to the placebo concurrent control in superiority trials. However, ethical use of a placebo concurrent group in assessment of efficacy and safety of new treatments in clinical trials has been continuously questioned and challenged; e.g., see Rothman and Michels (1994). As a result, it may not be ethical to conduct a placebo-controlled trial for serious illness or patients with severe or life-threatening diseases for establishment of the efficacy of a new drug product. On the other hand, as many drug products have been approved by regulatory agencies due to substantial evidence of efficacy and safety generated from placebo-controlled superiority trials for the treatment of a number of diseases, attention has focused on a search of new therapeutic modalities to compete with the standard and known effective drug products currently available on the market. These new products may offer some specific advantages over the standard drugs. These advantages include a better safety profile such as reduction of risk of intracranial hemorrhage, or of some grade 3 or 4 toxicities in some cancer treatments, an easy administration route such as from IV infusion to oral administration, a short duration of treatment, an improvement on quality of life, and most importantly, reduction of cost. Due to these reasons, the use of active concurrent control has recently become increasingly popular. However, despite their recent popularity, the use of active control trials is not without unresolved problems, issues, and challenges that have to do with the objectives of the active control trials.

Primary Objectives

For the development of a test drug product, the primary objectives of an active control trial could be

- 1.** To establish the efficacy of the test drug product.
- 2.** To demonstrate that the test drug product is superior to the active control agent.
- 3.** To show that the test drug product is similar to an active control agent or
- 4.** To verify that the test drug product is no worse than the active control.

Because the placebo concurrent control is not included in the two-arm active control trials, the efficacy of the test drug product cannot be established due to lack of direct evidence of effectiveness of the test drug product against placebo (Pledger and Hall, 1986; Temple and Ellenberg, 2000; Ellenberg and Temple, 2000). Although the hypotheses in (7.3.2) for superiority trials can be used to assess objective (2) for demonstration of superior efficacy of the test drug product to the active drug product, it still cannot establish the efficacy of the test drug product because the two-arm active control trials do not include a placebo concurrent control. Without inclusion of a placebo concurrent control in a two-arm active control trial, the efficacy of a test drug product may be inferior to or indistinguishable from that of placebo even though it is superior to the active control.

Objective (3) can be achieved through a two-sided equivalence trial with the primary goal of showing that the response to two or more treatments differs by an amount that is

clinically unimportant (ICH E9, 1998). This objective can be usually verified regarding whether the true treatment difference is to lie between a lower and an upper equivalence limits of clinically acceptable differences. It follows that the hypotheses corresponding to objective (3) as interval hypothesis given in Section 2.7:

$$\begin{aligned} H_0: \mu_T - \mu_C &\geq U \quad \text{or} \quad \mu_T - \mu_C \leq L \\ \text{vs.} \quad H_a: L < \mu_T - \mu_C < U, \end{aligned} \quad (7.4.1)$$

where L and U are some prespecified clinically meaningful lower and upper equivalence limits.

Average bioequivalence trials based on pharmacokinetic profiles of plasma concentrations of the active ingredients required for approval of generic drug products are typical two-sided equivalence trials as mentioned in Section 2.7 (FDA, 2001a). However, when drugs are not absorbed and plasma concentrations of their active ingredients are negligible, two-sided equivalence trials are also undertaken to demonstrate a two-sided clinical equivalence of the generic drug product to the innovator's product.

On the other hand, because of other advantages provided by the active control, many active control trials are designed to show that the efficacy of a test drug product is not clinically inferior to that of the active comparator. A trial with objective (4) is referred to as a noninferiority trial. The corresponding hypotheses to a noninferiority trial is given as

$$\begin{aligned} H_0: \mu_T - \mu_C &\leq L \\ \text{vs.} \quad H_a: \mu_T - \mu_C &> L. \end{aligned} \quad (7.4.2)$$

ICH E9 (1998) indicates that statistical analysis for evaluation of equivalence/noninferiority is generally based on the use of a confidence interval. Two-sided equivalence between the test drug product and the active control is concluded if the lower and upper limit of the confidence interval for $\mu_T - \mu_C$ is completely contained within (L, U) . On the other hand, if the lower limit of the confidence interval for $\mu_T - \mu_C$ is greater than L , then noninferiority of the test drug product is concluded. Again, both two-sided equivalence trials and one-sided noninferiority trials cannot provide direct evidence of the effectiveness of the test drug product because they do not include a placebo concurrent group.

Issues in Active Control Equivalence Trials

Issues, difficulties, and controversies surrounding the active control equivalence trials (ACET) and the interpretation of their results have long been well recognized and extensively documented, e.g., Temple (1983, 1996, 1997), Makuch and Johnson (1989), Leber (1989), Senn (1993), Jones et al. (1996), Ware and Antman (1997), Ebbutt and Firth (1998), ICH E9 guideline (1998), ICH E10 guideline (1999), Fisher et al. (2001), Fleming (2000), Siegel (2000), Temple and Ellenberg (2000), Ellenberg and Temple (2000), Djulbegovic and Clarke (2001), and Chow and Shao (2002a, 2002b).

On the other hand, Kirshner (1991) proposed a set of criteria for evaluation of the quality of active control equivalence trials that is reproduced in Table 7.4.1. The number one problem of any two-armed clinical control trials without a placebo concurrent control is the verification of the fundamental assumption of active control equivalence trials—the effectiveness of the active concurrent control. Had this fundamental assumption been true, then the effectiveness of the test drug product could have been established if it is shown to

Table 7.4.1 Criteria for Evaluation of Active Control Trials

-
1. Is the standard therapy effective, and does the experimental therapy have an advantage that is generalizable?
 2. Can an acceptable minimum effect be defined in terms of the outcome measures?
 3. Can this effect be measured precisely?
 4. Was the assignment of patients to treatments really randomized?
 5. Was the time frame for follow-up sufficient to conclude that the therapies are equivalent?
 6. Were factors that determined participation in the study identical to those for receiving the standard therapy?
 7. Is it probable that the groups differ by less than a minimum effect?
 8. Were all patients who entered the study accounted for at its conclusion?
 9. Were the treatments administered as they would be in practice?
-

Source: Kirshner (1991).

be equivalent to or not inferior to the active concurrent control. However, without a placebo concurrent control, the fundamental assumption of active control trials cannot be proven directly by the present active control trial. However, if this fundamental assumption is not true and cannot be verified, when the test drug product is equivalent to or not inferior to the active control, they can be both efficacious and both ineffectual.

Example 7.4.1 Active Control Trials of Antidepressants

Temple (1983, 2000) and Leber (1989) provided an excellent illustration of the danger for inference of the results from active control trials. A series of six trials was conducted to investigate the effectiveness of nomifensine (a test antidepressant) for treatment of endogenous depression with an active control imipramine, considered as a standard tricyclic antidepressant. The common baseline means and four-week adjusted group means based on the Hamilton depression scales were given in Table 2.7.1. Table 7.4.2 provides the mean changes at the 4 weeks from baseline and *p*-values and power for detecting a difference between nomifensine and imipramine. As Table 7.4.2 indicates, the sample sizes of these six trials range from 15 to 59. If we only consider the comparison between nomifensine and imipramine, each trial showed that both nomifensine and imipramine provided considerable improvement in Hamilton depression scale after four weeks of treatment. The two active drugs were in fact indistinguishable (*p*-value ≥ 0.63). The six trials seem to provide substantial evidence on the efficacy of nomifensine. Note that although there is low power for detection of a 30% difference (ranging from 0.09 to 0.45), the results also seem to be highly supportive of the effectiveness of both nomifensine and imipramine because the observed differences were greater than 30% from baseline that is considered a clinically meaningful difference.

Fortunately, all six trials also include a placebo concurrent control so that assay sensitivity of the trials can be evaluated. A three-way comparison revealed that there are no significant differences among placebo, nomifensine, and imipramine except for one study [V311(2)] that shows a detectable placebo-active drug difference. The results suggest that the historical assumption about imipramine may be erroneous. In other words, the sensitivity-to-drug effect of imipramine cannot be verified. This example argues against the idea that active control studies of putative antidepressants can serve as a sole basis for concluding that a drug is effective.

Table 7.4.2 Hamilton Depression Scale Endpoints—Mean Change Scores from Baseline

Study	Baseline	Placebo	Nomifensine	Imipramine	p-value ¹	Power ²
R301	23.9	-9.1(n = 36)	-10.5(n = 33)	-11.1(n = 33)	0.78	0.40
G305	26.0	-12.1(n = 36)	-13.0(n = 39)	-12.6(n = 30)	0.86	0.45
C311(1)	28.1	-9.2(n = 13)	-8.7(n = 11)	-7.8(n = 11)	0.81	0.18
V311(2)	29.6	-6.1(n = 7)	-22.3 ³ (n = 7)	-20.1 ³ (n = 8)	0.63	0.09
F313	37.6	-15.6(n = 8)	-15.7(n = 7)	-15.7(n = 8)	1.00	0.26
K317	26.1	-15.6(n = 36)	-14.9(n = 37)	-15.3(n = 32)	0.85	0.33

1. Two-tailed p-values for detection of a difference between nomifensine and imipramine.

2. Calculated for a 30% difference between nomifensine and imipramine.

3. Statistically significant difference from placebo at $p < 0.001$, two-tailed.

Source: Temple (1983) and Leber (1989).

Example 7.4.2 Trials of Beta-Adrenergic Blocking Agents

Temple (1983) indicates that antihypertensive drugs may represent a reasonable class which active control trials might be appropriate. These antihypertensive drugs regularly can be shown to be superior to the placebo in a fairly small trial. Table 7.4.3 shows that the results of five antihypertensive drugs, including metoprolol, pindolol, atenolol, nadolol, and timolol. It appears that these drugs are all different from the placebo, and all trials showed statistically significant drug-placebo difference. However, note that the placebo effect can be large in some trials. For example, in the metoprolol study, the placebo response of -11/6 mmHg was about as large as the drug response in the timolol study and better than the drug response in the nadolol trial or the pindolol study 25.

From these two examples, one critical issue of any two-arm active control trial is that there is no measure of internal validity for assay sensitivity of the trials. As shown in

Table 7.4.3 Results of Five Antihypertensive Trials of Marketed Beta-Adrenergic Blocking Agents Change in Blood Pressure in mmHg (Post/Baseline)

Study	N (Drug/Placebo)	Drug	Placebo	D – P
Metoprolol				
Reeves	11/12	-17/11	0/0	-17/11
Bowen	18/16	-19/14	-11/6	-8/8
Pindolol				
Study 25	29/28	-9/7	-2/3	-7/4
Study 24	28/30	-14/11	-5/4	-9/7
Atenolol				
Curry	15/12	-17/12	-5/4	-12/8
Nadolol				
Multicenter	50/50	-9/10	-3/2	-6/8
Timolol				
Multicenter	120/144	-11/10	1/2	-12/8

Source: Temple (1983).

Example 7.4.1, except for Study V311(2), inclusion of a placebo concurrent control was able to show that all other five trials lack assay sensitivity of an effect of imipramine of a specific clinically meaningful size as compared to the placebo group and that nomifensine is ineffective. Without assay sensitivity of imipramine, the results of these five trials will mislead us to the wrong conclusion of equivalence or noninferiority of nomifensine to imipramine when nomifensine is in fact inferior. On the other hand, because the actual effect size of the active control is not measured in two-arm active control trials, the presence of assay sensitivity must be deduced and external validation for the effectiveness of the active comparators becomes necessary. External validation of sensitivities-to-drug-effects is determined from historical experience of the active control.

Validation of historical evidence of sensitivity-to-drug-effects for the active control is extremely challenging and difficult. As pointed out by ICH E10 (1999), there are many circumstances in which the effectiveness of drugs considered efficacious cannot regularly be reproduced in well-controlled studies. These conditions include those in which substantial improvement and variability exists in placebo concurrent groups, and/or in which the effect of the active control is small or variable. An excellent illustration is given in Example 7.4.2. There is a large variation of change in blood pressure in both beta-blocker groups and placebo. Changes in blood pressure for the placebo groups in seven trials range from -11 mmHg to 1 mmHg . In addition, the reduction in blood pressure in the placebo group of -11 mmHg of the metoprolol-Bowen study was better than that of the active beta-blockers in three out of the seven trials shown in Table 7.4.3. Other indications with similar problems are depression, anxiety, dementia, angina, symptomatic congestive heart failure, seasonal allergies, and symptomatic gastroesophageal reflux disease. In conclusion, even though there is no doubt that the active control is indeed effective because it has been proven in many well-controlled studies, it is still quite difficult to describe study conditions in which the active control would reliably have a minimum effect for determination of equivalence limits (ICH E10, 1999). In addition, the historical experience of sensitivity-to-drug-effects also evolves over time as the diagnosis and standard treatment also change rapidly over time. Therefore, if an ACET cannot provide acceptable support for the sensitivity-to-drug-effects with the selected equivalence limits, a finding of equivalence/noninferiority cannot be considered informative with respect to efficacy or to showing of clinical equivalence.

Even though the historical evidence indicates that the proposed ACET in a particular indication is likely to have sensitivity-to-drug-effects, this likelihood can be jeopardized by the poor conduct of the trials. ICH E10 guideline lists the following factors that can reduce ACET assay sensitivity:

- 1.** Poor compliance with therapy
- 2.** Poor responsiveness of the study population to the active control
- 3.** Use of concomitant medication or other treatment that interferes with the test drug product or that reduces the extent of the potential response
- 4.** A population that tends to improve spontaneously, leaving no room for further drug-induced improvement
- 5.** Poor diagnostic criteria
- 6.** Inappropriate or insensitive measures of drug effects
- 7.** Excessive variability of measurements
- 8.** Biased assessment of endpoints

Use of the techniques of randomization and blinding as described in Chapter 4 can in general improve the assay sensitivity and reduce the bias associated with investigators and subjects. However, an obvious difference between the ACET and placebo-controlled trial is that both investigators and subjects know that all subjects in the ACET receive a potential active drug. This particular knowledge may lead to bias in evaluation and measurement of endpoints, and interpretation of the results. For example, an investigator may tend to read blood pressure responses greater than they actually are. This will reduce the difference between the test drug product and the active control. Another example is a biased interpretation of the results in the form of a tendency toward categorizing borderline cases as success in partially subjective evaluations, e.g., in antidepressant trials. Such biased evaluation and interpretation of the results may reduce variability and/or decrease treatment differences and hence increase the likelihood of incorrect conclusion of equivalence.

To ensure sensitivity-to-drug-effects of an equivalence/noninferiority trial, the ICH E9 guideline (1998) suggest that a suitable active control be a widely used therapy whose efficacy in the relevant indication has been clearly established and quantified in well-designed and well-documented superiority trials and that can be reliably expected to show similar efficacy in the intended active control equivalence trial. Therefore, the design and conduct of the new ACET should be similar to placebo-controlled studies of the active control in the past that successfully demonstrated the effectiveness of the active control. Close attention should be paid to important design characteristics such as the same inclusion/exclusion criteria for the study population (severity of medical condition, method of diagnosis), the same design, primary endpoints measured and timing and method of assessment, dose and dosing schedules of the active control, and the use of washout or run-in period to exclude patients without disease or with spontaneous improvement. In addition, the study conduct of the trial should be also closely monitored to ensure the study's assay sensitivity. These aspects of study conduct include compliance, monitoring of concomitant medications, enforcement of adherence to inclusion/exclusion criteria, and prevention of study dropout. Furthermore, the design and analysis of any ACET should also take into account advances in medical or statistical practice relevant to the intended active control equivalence trial. More about the issues in active control trials can be also found in Hwang (2001), Hung (2001), Snapinn (2001), Tsong et al. (2001), Wittes (2001), Kallen and Larsson (2001), and Chow and Shao (2002b).

Interpretation of the Results of Active Control Trials

The interpretation of the results of active control trials regarding the efficacy and safety of the drug product under investigation is extremely important in clinical research. In what follows, examples concerning active control trials are presented to illustrate how the results should be interpreted.

Example 7.4.3 AIDS Antiretroviral Equivalence Trials

The objective of an antiretroviral therapy is to achieve prolonged suppression of human immunodeficiency virus (HIV) replication. A standard approach to initial antiretroviral treatment is with two nucleoside analogues and a protease inhibitor. Although protease containing antiretroviral regimens can accomplish the goal of delaying progression of the acquired immunodeficiency syndrome (AIDS) and increase survival, they are not without problems that limit their long-term effectiveness and complete viral suppression. These

problems include poor tolerability, metabolic toxic effects, and drug interaction due to inhibition of cytochrome P450 enzymes and poor compliance due to the complexity of dosing regimens. Example 3.4.6 indicated that one of the conventional antiretroviral therapies for the treatment of previously untreated HIV infected patients is a triple nucleoside analogue regimen of indinavir-lamivudine-zidovudine (the ILZ regimen). This particular regimen asks the patients to take 150 mg of lamivudine and 300 mg of zidovudine twice daily and 800 mg of indinavir in 200-mg capsule formulation every 8 hours daily for a total of 16 tablets per day. In addition, indinavir is administrated with water 1 hour before or 2 hours after a meal and patients should drink at least 1.5 liter of water during the course of 24 hours to avoid dehydration. As a result, this triple nucleoside analogue regimen of indinavir-lamivudine-zidovudine poses considerable inconvenience to the patients and adherence of the regimen by patients.

Abacavir is a potent inhibitor of HIV reverse transcriptase (TR) that dose does not select resistant virus *in vitro*. In addition to its antiretroviral activity as monotherapy, marked antiretroviral activity has also been demonstrated in combination of lamivudine-zidovudine. In addition, the triple nucleoside analogue regimen of abacavir-lamivudine-zidovudine (the ALZ regimen) only requires patients to take a 300-mg tablet of abacavir twice daily for a total of four tablets per day. In lieu of several advantages of the ALZ regimen such as twice daily dosing, low pill burden, low drug interaction, and the potential to reserve other drug classes for future treatment options, the CNAAB3005 International Study Team (Staszewski et al., 2001) conducted the first randomized, double-blinded trial to evaluate the two-sided antiretroviral equivalence between the ALZ and ILZ regimens for initial treatment in antiretroviral-naïve HIV-infected adults. Primary endpoints were based on the proportion of the patients with the reduction of plasma HIV RNA copies/mL below some prespecified level after 48 weeks of treatment. Four hundred copies and 50 copies/mL were the two levels considered in this trial. After discussion with investigators and with the US FDA, a symmetric lower and upper equivalence limit of $\pm 12\%$ were prespecified as the largest difference that would be considered as clinically acceptable. The two triple nucleoside analogue regimens of ALZ and ILZ were considered equivalent if the 95% confidence interval for the difference between the regimens is within the limits from -12% to 12% . From these considerations, the sample size of 275 patients per group provided sufficient power to achieve the objective of two-sided equivalence.

Table 7.4.4 presents a summary of patient enrollment and discontinuation of this trial, reproduced from Figure 1 in Staszewski et al. (2001). Seven hundred and eighty one subjects were screened, and 562 patients were randomized: 282 to ALZ group and 280 to ILZ group. Twenty patients in ALZ group and 15 patients in ILZ group did not take the study drugs at the start of the trial. Overall, 99 patients in ALZ group and 96 patients in ILZ group discontinued the study prematurely for various reasons. A total of 316 patients completed a 48-week randomized treatment: 150 in ALZ group and 156 in ILZ group. As a result, a total of 527 patients were included in the population of the intention-to-treat analysis that consists of all randomized patients with any post-treatment measurements. However, only between 284 and 297 patients were included in the population of as-treated analysis that consists only of the patients continuing randomized treatment. As a result, 43.6% to 46.1% of patients in the intention-to-treat analysis were not included in the as-treated analysis. Note that in Table 7.4.4, it is not clear from Figure 1 in Staszewski et al. (2001) why the sum of the numbers of patients not receiving the treatment, switching treatment, discontinuing study, and completing the study is not added to the number of patients randomized.

Table 7.4.4 Summary of Patient Enrollment and Discontinuation for CNAAB3005 International Study Through 48 Weeks of Treatment

Event	Total	ALZ Group	ILZ Group
Screening	781		
Excluded	219		
Randomization	562	282	280
Not receive treatment	35	20	15
Switch of treatment	28	11	17
Discontinuation	195	99	96
Completing the study	316	160	156
Intention-to-treat	527	262	265
As treated	284–297	145–150	139–147

Source: Staszewski et al. (2001).

Table 7.4.5 presents the point estimates and their corresponding 95% confidence intervals for the treatment difference on the proportion of patients with a plasma HIV RNA level either no greater than 400 or 50 copies/mL. The primary efficacy endpoint preselected by the CNAAB3005 International Study Team was the proportion of patients with plasma HIV RNA level no greater than 400 copies/mL. For the intention-to-treat analysis, at week 48, 51% of the patients (133/262) in ALZ group and 51% (136/265) in ILZ group had sustained suppression of HIV RNA levels to less than 400 copies/mL. The treatment difference is estimated as 0.6% with the corresponding 95% confidence interval of (−9%, 8%). Because the 95% confidence interval of (−9, 8%) is completely contained within the equivalence limits of (−12%, 12%), the CNAAB3005 International Study Team concluded that the efficacy of the triple nucleoside analogue regimens of ALZ based on the primary endpoint is equivalent to that of IZT (Staszewski et al., 2001). However, the as-treated analysis yielded a treatment difference of −7% with the corresponding 95% confidence interval of (−14%, 0%). Therefore, with respect to the population of as-treated analysis, equivalence between the ALZ and ILZ cannot be concluded because the 95% confidence interval is not contained within

Table 7.4.5 Summary of Results on the Difference in the Proportion of the Patients with a Plasma HIV RNA Level no greater than 400 or 50 copies/mL for CNAAB3005 International Study

Endpoint	Population	Point Estimate	95% Confidence Interval
% ≤ 400	Intention-to-treat	0.6%	(−9%, 8%)
	As treated	−7%	(−14%, 0%)
% ≤ 50	Intention-to-treat	−6%	(−15%, 2%)
	As treated	−13%	(−23%, −4%)
	Intention-to-treat and high HIV RNA baseline	−14%	(−27%, 0%)
	Intention-to-treat and low HIV RNA baseline	−2%	(−13%, 9%)
	As treated and high HIV RNA baseline	−12%	(−23%, −1%)
	As treated and low HIV RNA baseline	−17%	(−34%, 0%)

Source: Summarized from Staszewski et al. (2001).

(−12%, 12%). Furthermore, one can reach the conclusion that, based on the as-treated analysis for this primary endpoint, the efficacy of the ALZ group in fact is inferior to that of ILZ group. If one uses a more stringent criterion for sustained suppression of HIV RNA level as 50 copies/mL, it can be shown in Table 7.4.5 that neither the intention-to-treat and as-treated analyses reach the conclusion of equivalence between the ALZ and ILZ groups. As a matter of fact, based on this criterion, the as-treated analysis yielded a finding that the efficacy of ILZ is superior to that of ALZ. Other conflicting findings can be also found for patients with low or high HIV RNA baseline (10,000–100,000 copies/mL or >100,000 copies/mL) in Table 7.4.5. The inconsistency of the results between the intention-to-treat and as-treated analyses may be due to more than 41% of patients who discontinued the study prematurely before its schedule completion of randomized treatment at week 48. Therefore, Djulbegovic and Clarke (2001) commented that the apparent equivalence reflected in the intention-to-treat analysis might simply be due to a dilutional effect of comparing the patients between ALZ and ILZ groups whose actual treatments did not differ much. In addition, the ICH E9 guideline also indicates that the noncompliance in the intention-to-treat analysis will generally diminish the estimated treatment effect. As a result, in equivalence or noninferiority trials, unlike the superiority trials, the use of the intention-to-treat analysis is not conservative and its role should be considered very carefully.

Equivalence/Noninferiority Limits

In addition to the issue of analysis sets, the determination of equivalence limits is also critical and yet extremely controversial. For Example 7.4.3, it is not clear how the equivalence limit of (−12%, 12%) was derived, and its clinical interpretation, relevancy, and implication are also vague. Another issue for selection of equivalence limits is whether the same equivalence limits should be used for different endpoints. For example, should the same equivalence limits of (−12%, 12%) be used for the cutoff points of 400 copies/mL and 50 copies/mL for sustained suppression of HIV RNA levels? The criterion of 400 copies is more lenient than that of 50 copies. In addition, different PCR assays might have a different limit of quantification that may lead to selection of different equivalence limits. For the CNAAB3005 International Study, two different PCR assays with limits of quantification of 400 copies/mL and 50 copies/mL were used. One final note is that 284 patients were included in the as-treated analysis using the criterion of 400 copies. On the other hand, for more stringent criterion of 50 copies/mL, 297 patients were included in the as-treated analysis. This inconsistency in the number of patients in the two as-treated analyses might be due to the use of different PCR assays, as described above.

Example 7.4.4 Continuous Infusion versus Double-bolus Administration of AMI

Example 3.4.5 indicated that the objective of the Continuous Infusion versus Double-bolus Administration of Alteplase trial (COBALT) investigators (1997), is to test the hypothesis that double-bolus alteplase is at least as effective as accelerated infusion. The primary endpoint for assessment of noninferiority in double-bolus alteplase in treatment of the patients with acute myocardial infarction was death from any cause at 30 days. This study employed the one-sided noninferiority hypothesis to evaluate the clinical equivalence between two administrations of alteplase. The double-bolus group was considered not inferior to accelerated infusion if the one-sided 95% upper confidence limit for the difference on 30-day mortality between the double-bolus administration and accelerated infusion is less than

0.4%. A total of 7,169 patients with acute MI were randomly assigned to receive accelerated infusion (3,585) or to receive the double-bolus administration (3,584). The COBALT investigators reported that 30-day mortality was higher in the double-bolus group than in the accelerated infusion group: 7.98% as compared with 7.53%. The point estimate for treatment difference in 30-day mortality between the double-bolus and accelerated infusion group is 0.44% with the corresponding one-sided 95% upper limit of 1.49%. As the one-sided 95% upper limit is 1.49%, which is greater than the prespecified noninferiority limit of 0.4%, despite the advantage of the ease of use, the evidence from the COBALT study failed to support that double-bolus alteplase is equivalent to accelerated infusion with respect to 30-day mortality.

As indicated earlier, the selection of equivalence critical limits is probably the most controversial and yet most critical in the design and analysis of ACETs. The ICH E10 guideline indicates that the noninferiority limits is the degree of inferiority of the test drug product compared to the active control that the ACET will exclude statistically. In addition, the limit selected for noninferiority trial cannot be greater than the smallest effective size that the active control would reliably expect to have compared with the placebo. If a difference between the test drug product and the active control favors the active control by as much as or more than that magnitude, the test drug product might have no effect at all. The equivalence/noninferiority limits are usually chosen based on past experience and historical evidence in placebo-controlled trials of adequate design similar to intended ACETs. Determination of equivalence/noninferiority limits is based on both clinical judgment and statistical reasoning that should reflect uncertainties in the evidence and it should be conservative.

For fibrinolytic therapy in treatment of suspected acute myocardial infarction, the 30-day mortality rate is around about 12% for placebo group and is about 8% for tissue plasminogen activator (t-PA, or alteplase) (Collins et al., 1997). As a result, the treatment effect of t-PA against placebo is estimated as 4%. Example 7.4.4 indicated that the COBALT investigators (1997) employed an equivalence limit of 0.4%. This limit is one-tenth of the estimated relative treatment effect against placebo. This implies that the upper limit of deaths allowed for the double-bolus alteplase to be considered therapeutic equivalent to accelerated infusion is 4 more deaths per 1,000 patients. This number is fewer than 5.6 per 1,000 patients between alteplase and streptokinase reported by Collins et al. (1997). This noninferiority limit is very conservative because it preserves 90% of the relative efficacy of t-PA against placebo. Therefore, Ware and Antman in their editorial (1997) emphasized that a sample size of 50,000 would be required to provide a 80% power to rule out excessive 30-day mortality rates of 0.4% when the true mortality rates are identical and in the range of 7.5%. Another remark is that the sample size of 7,169 patients for the COBALT study is to provide an 80% power to actually reduce a 30-day mortality rate from 6.3% to 5.4% for a superiority trial.

As another example, in treatment of perennial allergic rhinitis, the efficacy is assessed based on an average daily total symptom score over duration of treatment that is the sum of four individual symptom scores, each with a range from 0 to 3. It seems reasonable to select an equivalence limit of 3.2 for evaluation of therapeutic equivalence between a test drug product and the active control because 3.2 represents 25% of the range for the total symptom scores. However, this limit of 3.2 is considered unreasonably liberal if one realizes that improvement in each of the four individual symptom scores provided by the active control over placebo estimated from adequate well-controlled studies with sufficient power is only about 0.5.

From the above discussion, there are three possible noninferiority limits (margins) that could be used for evaluation of noninferiority hypothesis in (7.4.2) (Temple and Ellenberg, 2000). The first one denoted as L_1 is the smallest effect the active control can be presumed to have in the study compared to a placebo concurrent control. For fibrinolytic therapy in treatment of suspected acute myocardial infarction, as demonstrated before, L_1 is about -4% . The second one denoted as L_2 is a fraction of L_1 ; it is selected because it is considered important to assure that the test drug product retains a substantial fraction of the effect of the active control. For the COBALT study, L_2 is chosen as 0.4% that reflects retention of 90% of the effect of the accelerated infusion or that the maximally acceptable loss of the accelerated infusion with the double bolus is 10% . Another well-known example is the frequently cited *50% rule* depicted in Center of Biological Evaluation and Research, U.S. FDA documents (CBER/FDA, 1999). The last one denoted as L_3 is selected as 0 . The relationship among these three limits in noninferiority hypothesis in (7.4.2) can be described as $L = -f(\mu_C - \mu_P)$, where μ_C and μ_P are the means of active control and placebo control in the placebo-controlled studies conducted previously, assuming that they remain unchanged over time:

$$\begin{aligned} H_0: \mu_T - \mu_C &\leq -f(\mu_C - \mu_P) \\ \text{vs. } H_a: \mu_T - \mu_C &> -f(\mu_C - \mu_P). \end{aligned} \quad (7.4.3)$$

$f(\mu_C - \mu_P)$ in (7.4.3) represents the minimum fraction of the effectiveness of the active control relative to placebo preserved by the current ACET, given that the effectiveness of the active control relative to placebo has been established by previously conducted adequate placebo-controlled trials.

This set of hypotheses can be reformulated as

$$\begin{aligned} H_0: \mu_T - \mu_P &\leq (1-f)(\mu_C - \mu_P) \\ \text{vs. } H_a: \mu_T - \mu_P &> (1-f)(\mu_C - \mu_P). \end{aligned}$$

If $f = 1$, $L = L_1$ and the noninferiority hypothesis in (7.4.3) becomes a superiority hypothesis of the test drug product over the placebo:

$$H_0: \mu_T \leq \mu_P \quad \text{vs. } H_a: \mu_T > \mu_P.$$

When $f = 0$, $L = L_3$ and the noninferiority hypothesis in (7.4.2) becomes a superiority hypothesis of the test drug product over the active control:

$$H_0: \mu_T \leq \mu_C \quad \text{vs. } H_a: \mu_T > \mu_C.$$

Temple and Ellenberg (2000) emphasized that L_3 must be used when the active control is not regularly superior to placebo in which $f = 0$. As a result, in this situation, only superiority of the test drug product over the active control in the current ACET may be accepted as evidence of effectiveness, even though, as mentioned before, there is a possibility that both the test drug product and the active control could be worse than placebo. In most situations, the efficacy of the active control as compared to the placebo concurrent control has been demonstrated previously, and L_2 is the most common used equivalence/noninferiority limit. The fraction that L_2 must retain, however, depends on the smallest clinically important

difference and the degree of advantages that the test drug product can offer (Fleming, 2000). Wiens (2002) and Wang and Hung (2003) reviewed and proposed statistical methods for determination of equivalence limits that include criteria such as some fraction of variation or proportion of similar responses.

Another challenge in determination of equivalence limits is their clinical interpretation and consequent clinical implication. For fibrinolytic therapy in treatment of suspected acute myocardial infarction, the GUSTO I trial (GUSTO I, 1993) provided a strong evidence that the smallest clinically important difference in 30-day mortality rates should be less than 1% despite the fact that alteplase has a higher cost and higher intracranial hemorrhage (Fleming, 2000). A naive question is that, is a noninferiority limit of 0.4% employed by the COBALT investigators too stringent and should a more lenient equivalence limits, say, 0.75% or 1%, be used? The answer is no. Suppose that the double-bolus alteplase is clinically equivalent to the accelerated infusion based on a noninferiority limit of 0.75% for the difference in 30-day mortality rates. Another new easy administration of alteplase, say, formulation B, could be concluded to be no worse than the double-bolus alteplase using the same equivalence limit of 0.75%. Then, an even newer and easier administration of alteplase formulation C could be accepted as a noninferior therapy to formulation B, again using the same limit of 0.75%. It follows that formulation C can replace the accelerated infusion, and even formulation C provides a 2.25% higher 30-day mortality rate than the accelerated infusion. Chow and Liu (1997) and Fleming (2000) gave excellent illustrations of this phenomenon.

Statistical Methods

Example 7.4.5 GUSTO III Trial Reteplase (recombinant plasminogen activator) is a mutant of alteplase tissue plasminogen activator and has a longer half-life than its parent molecule. Therefore, it produced superior angiographic results in patients with acute myocardial infarction. GUSTO III investigators (GUSTO III, 1997), therefore, conduct a randomized trial to test the hypothesis that the double-bolus reteplase is superior to the accelerated infusion of alteplase in a 30-day mortality rate. A total of 15,059 patients were randomly assigned in a 2:1 ratio to receive double-bolus reteplase (10,138 patients) or accelerated infusion of alteplase (4,921 patients). The 30-day mortality rate was 7.47% for reteplase and 7.24% for alteplase. The point estimate for the difference in 30-day mortality rates is 0.23% with the corresponding 95% confidence interval from -0.66% to 1.1%. Because the 95% confidence interval includes 0 and, hence, there is no statistically significant difference in 30-day mortality rates, the GUSTO III investigators declared that the double-bolus reteplase and accelerated infusion of alteplase had similar clinical efficacy.

Because the upper limit of the 95% confidence interval is 1.1%, with the noninferiority limit of 0.4% used by the COBALT, the data of the GUSTO III do not demonstrate equivalence between reteplase and alteplase (Ware and Antman, 1997). One crucial mistake that the GUSTO III investigators made is to reach a conclusion of equivalence because there is a small and nonsignificant difference. The fundamental concept for equivalence is the concept that lack of evidence of a difference is not evidence of lack of a difference (Altman and Bland, 1995). Therefore, the ICH E9 guideline (1998) stressed that concluding equivalence or noninferiority based on observing a nonsignificant test result of the null hypothesis that there is no difference between the test drug product and the active control is inappropriate (also see Chow and Liu, 2000).

The most critical issue in analysis for evaluation of equivalence/noninferiority in a two-arm active control trial is lack of a placebo concurrent control. The U.S. FDA requests that the effectiveness of a test drug product must be established by comparisons with the placebo concurrent control. However, because the two-arm active control trial does not include a placebo concurrent control, the effectiveness of the test drug is established through comparisons with a placebo nonconcurrent (putative or ghost) control in the hypothesis formulation for a superiority trial (Tsong et al., 2001):

$$\begin{aligned} H_0: \mu_T - \mu_{PN} &\leq 0 \\ \text{vs. } H_a: \mu_T - \mu_{PN} &> 0, \end{aligned} \quad (7.4.4)$$

where μ_{PN} is the mean of the placebo putative control.

Note that because a two-arm active control trial does not include a placebo concurrent control, μ_{PN} cannot be estimated from the current ACET. Tsong et al. (2001) proposed the following equation to explore the relationship between $\mu_T - \mu_{PN}$:

$$\mu_T - \mu_{PN} = (\mu_T - \mu_C) + (\mu_C - \mu_{CH}) + (\mu_{CH} - \mu_{PH}) + (\mu_{PH} - \mu_{PN}), \quad (7.4.5)$$

where μ_C is the mean of the active concurrent control, μ_{CH} is the mean of the active historical control, and μ_{PH} is the mean of the placebo historical control.

If $\mu_C - \mu_{PN} = \mu_{CH} - \mu_{PH}$, then (7.4.5) can be reexpressed as

$$\mu_T - \mu_{PN} = (\mu_T - \mu_C) + (\mu_{CH} - \mu_{PH}). \quad (7.4.6)$$

The assumption that the efficacy of the active concurrent control with respect to the placebo putative control is the same as the efficacy of the active historical control relative to the placebo historical control is called the constancy assumption of the effective size of the active control. In other words, the effect of the active control to placebo is assumed unchanged from historical placebo-controlled trials to the current ACET. This assumption is extremely difficult to verify. Under the constancy assumption, hypothesis of superior effectiveness in (7.4.4) can be reformulated as

$$\begin{aligned} H_0: \mu_T - \mu_C &\leq -(\mu_{CH} - \mu_{PH}) \\ \text{vs. } H_a: \mu_T - \mu_C &> -(\mu_{CH} - \mu_{PH}). \end{aligned} \quad (7.4.7)$$

In hypotheses (7.4.7), $\mu_T - \mu_C$ on the left-hand side of hypotheses is the difference in between the test drug product and the active concurrent control in the current ACET. On the other hand, $\mu_{CH} - \mu_{PH}$ on the right-hand side of (7.4.7) is the effective size of the active control from the historical placebo-control trial. A comparison of hypothesis (7.4.7) and (7.4.2) reveals that to show the effectiveness of the test drug product, the equivalence limit is determined by the effective size of the active control from the historical placebo-control trial. In the current ACET, patients are randomized to either the test drug product or to the active control but not to the “putative” placebo. One of the consequences is that randomization-based inference is not possible to perform hypothesis in (7.4.4). Although within-trial randomization can minimize the bias in estimation of the parameters within historical trials and current ACET, bias in comparison of the parameters across the trials still exists.

Several different statistical methods are currently available in literature for evaluation of the effectiveness of the test drug products. The first method is called the indirect confidence interval comparison (ICIC) that entails prespecification of the noninferiority limit against which comparison of the test drug product with the active control. The ICIC method uses the equivalence/noninferiority limit L_2 described above to preserve at least a certain fraction of the effectiveness of the active control relative to the placebo by the test drug product. Once the equivalence/noninferiority limit is determined, the test drug product is concluded efficacious if the lower $(1 - \alpha)100\%$ confidence limit is greater than L_2 . The other method is referred to as the virtual comparison (VC) that synthesizes the estimated relative efficacy of the test drug product to placebo from the current ACET and the relative efficacy of the active control to placebo from the historical placebo-controlled trials. Then the resulting statistics are treated as if it were from the same trial and the null hypothesis in (7.4.3) is performed by the traditional statistical methods such as t-test or z-score. Note that unlike the ICIC method, the VC method does not use the equivalence/noninferiority margin. In addition, the VC method can also provide an estimate of the fraction of the effectiveness of the active control retained by the test drug product (Hasselbald and Kong, 2001).

As mentioned before, both methods are based on the constancy assumption and a reasonable estimation from the historical placebo-control trials on the effectiveness of the active control relative to placebo. The random-effect model through normal approximation proposed by DerSimonian and Laird (1986) is often used to estimate the relative effectiveness of the active control from a collection of historical trials. Wang et al. (2002) reported empirical evidence of type I error rate (false positive) and power (true positive) from an extensive simulation study comparing these two methods. They found that the VC method is optimal under the constancy assumption. This method is not valid and cannot control the false-positive rate at the some prespecified level if the constancy assumption is violated in the sense that the effectiveness of the active control in the current ACET is smaller than that from the historical trials. They also reported that the VC method is not valid even under the constancy method when the effectiveness of the active control is estimated from a small number of historical trials by the normal approximation in the random-effects model proposed by DerSimonian and Laird (1986). In addition, they found that the method proposed by Hasselbald and Kong (2001) for estimation of the fraction of the effectiveness of the active control preserved by the test drug product is also misleading because of an inflated false-positive rate. In contrast, their simulation suggests that the ICIC method is ultraconservative in terms of a false-positive rate when the assumptions are met. However, their simulation also shows that the ICIC method can be liberal when the estimated efficacy of the active control in the current ACET is substantially less than that from the historical trials. To overcome the shortcomings of the ICIC and VC methods, Wang and Hung (2003) proposed the two-stage active control testing (TACT) method for evaluation of noninferiority in active control trials. Their method consists of three steps:

- Step 1: Validation of sensitivities-to-drug-effects for the active control
- Step 2: Verification of the constancy assumption of the active control
- Step 3: Establishment of effectiveness of the test drug product through noninferiority testing

Simulation studies suggest that TACT perform better than the ICIC and VC methods in terms of controlling the false-positive rate. Concepts and other statistical methods for

noninferiority trials can be found in a recently published special issue on noninferiority trial in *Statistics in Medicine* (D'Agostino, 2003).

7.5 DOSE-RESPONSE TRIALS

All drugs are potent and poisonous. Only at the right doses, the benefit provided by the drug exceeds its inherent risk and only then the drug is useful. Therefore, Moore (1995) pointed that the difference between a drug and a poison is the dose. Information of the relationship among dose, drug concentration, and clinical responses is extremely important for the safe and effective use of the drug. A common wrong impression about the dose-response information is only aimed at the relationship on efficacy or benefit. However, it should be emphasized that the dose-response relationship on safety is as equally important as to that on efficacy, if not more important. Critical dose-response information on both efficacy and safety can help identify an appropriate starting dose, the proper way to adjust dosage to the need of a particular patient, and a dose beyond which increases would be unlikely to provide additional benefit or would induce unacceptable adverse events (ICH E4, 1994). Determination of the dose and dosing range is the most difficult and challenging task during the clinical development of a test drug product. The issue of whether the dose-response relationship has been sufficiently characterized is one of the frequently asked questions at the U.S. FDA advisory committee. Tables 3.4.2 and 3.4.4 in Chapter 3 provided two excellent examples on failure of characterization for dose-response relationship for two test drug products in two different therapeutic areas. Trials conducted to characterize the dose-response relationship of the test drug product are referred to as dose-response trials.

The following objectives of dose-response trials are given by the ICH E9 guideline (1998):

1. Confirmation of efficacy.
2. Investigation of the shape and location of the dose-response curve.
3. Estimation of an appropriate starting dose.
4. Identification of optimal strategies for individual dose adjustments.
5. Determination of a maximal dose, beyond which additional benefit would unlikely to occur.

From the above objectives of dose-response trials, it is extremely important to identify the dosing range of a test drug product and to describe the relationship of the dose and clinical efficacy and safety responses within that range. Therefore, there are at least two types of dose-response curves with respect to the responses. The first is the dose-efficacy curve that describes the relationship between dose and some primary efficacy endpoints. The other is the dose-safety curve that characterizes the relationship between dose and safety endpoints, such as incidence of some adverse event or changes in some important laboratory evaluations. The dose-response curve can also be classified as the population average dose-response curve (PDRC) and individual dose-response curve (IDRC). The PDRC is useful for description of the shape and location of the response curve based on the average for the entire patient population. On the other hand, the IDRC is extremely important for selection of an initial starting dose and subsequent dose adjustment for an individual patient because many factors can contribute to differences in pharmacokinetics of the test drug products

among patients, including demographic factors such as age, gender, race; concurrent illness such as renal or hepatic failure; diet, concomitant therapy, or characteristics of individual patients such as weight or genetic differences.

The dosing range or therapeutic window is determined jointly by the dose-efficacy curve and the dose-safety curve. The lower bound of the dosing range is referred to as the minimum effective dose (MED). The MED is then defined as the lowest dose level of a test drug product that provides a clinically significant response in average efficacy that is also statistically significantly superior to the response provided by the placebo (Rodda et al., 1988; Ruberg, 1995a). In addition, the MED should also present a safety profile that is no worse than the placebo. According to this definition, the MED must produce a response with a magnitude of clinical superiority over placebo because a small but statistically significant response resulting from either large sample sizes or small variability can be of no clinical application. On the other hand, the MED must also provide a statistically significant clinical response. This is because a large clinical response, yet statistically insignificant from the placebo response produced at a dose level, cannot establish the scientific evidence of the effectiveness at that dose level. Similarly, the maximum tolerable dose (MTD) is the highest possible but still tolerable dose level with respect to a prespecified clinical limiting toxicity (Storer, 1989; Korn et al., 1994). The maximum useful dose (MUE) is referred to as the highest dose beyond which increases would unlikely provide additional efficacy. The dosing range or therapeutic window is then defined as the range of the dose levels from the MED to the minimum of MTD and MUE. If the difference between the maximum of MTD and MUE and MED is large, then the corresponding test drug product is said to have a wide therapeutic range. On the other hand, if the maximum of MTD and MUE is very close to or even smaller than the MED, then the product is practically of little therapeutic value.

The dose-response curve is usually estimated based on the data of primary efficacy endpoints collected from dose-ranging or dose-response clinical trials conducted during phase II clinical development of a drug product. These dose-response studies are usually randomized, double-blind, parallel-group designs comparing several doses to a concurrent placebo. Occasionally, a crossover design such as William's design (Chow and Liu, 2000) is employed. Many clinicians, however, find a variety of titration designs either with or without a concurrent parallel placebo group for dose-ranging studies that are useful due to the impression that they mimic the clinical practices in the real world. The ICH E4 guideline (ICH, 1994) classifies the dose-response designs as (1) parallel does-response design, (2) crossover dose-response design, (3) forced titration design, and (4) optional titration design (placebo-controlled titration to endpoints).

Randomized Parallel Dose-Response Designs

Randomized parallel dose-response design (RPDRD) is a straightforward application of the parallel group design described here to investigate the dose-response relationship. It is simple in concept and easy to implement with straightforward analyses and interpretation of the results. RPDRD has had extensive use and considerable success in obtaining and characterizing the dose-response relationship. For RPDRD, patients are randomly assigned to receive one of several fixed-dose groups. The fixed dose here is referred to as the final or maintenance dose. Patients in RPDRD may initially be placed on a starting dose and gradually titrated to the final dose according to a prespecified schedule. The study period of the final dose should be long enough for the full effect of the test drug product to realize and to

allow dose-response comparison. However, PRDRD provides only the population average dose response but not the distribution or shape of the individual dose-response curve.

The ICH E4 guideline indicates that it is not necessary to include in placebo concurrent control in RPDRD, and a positive slope provides evidence of a drug effect. However, it is highly recommended that a placebo concurrent control must be included in RPDRD to establish the effectiveness of the test drug product. A placebo concurrent control can avoid studies that are uninterpretable because all doses produce similar effects so that one cannot evaluate whether all doses are equally effective or equally ineffective. In addition, by inclusion of a placebo and active concurrent controls, one allows assessment of assay sensitivity, permitting a distinction between an ineffective dose and an ineffective trial, as shown in Tables 3.4.2 and 3.4.4. Furthermore, to measure the absolute size of the drug effect, a placebo concurrent control or an active concurrent control with very limited effect on the endpoint of interest must be included in RPDRD. The ICH E4 guideline also indicates that there is no need to detect a statistically significant difference in pairwise comparisons between doses if a statistically significant positive slope across doses can be established using all data. However, if the lowest dose employed in the trial is recommended as the MED, it should demonstrate a statistically significant and clinically meaningful effect. The primary objective of RPDRD is to establish the effectiveness of the test drug product and to characterize the dose-response relationship. Because detection of a positive slope requires fewer subjects than does detection of a difference and the drug-placebo difference is in general larger than the between-dose difference, RPDRD allow smaller sample sizes than do the pivotal phase III trials. In addition, unequal allocation of sample sizes can be employed in RPDRD design to allow more subjects assigned to the proposed MED and placebo group for obtaining more precise information and increasing power about the effect of the MED.

Objective (3) for dose-response trials is to estimate an appropriate starting dose, in other words, to estimate MED. To choose an appropriate statistical design, the selection of dose levels, the number of dose levels, and sample sizes for each dose group are equally important for estimation of the MED. These issues are not only related to each other, but also very difficult to deal with. The dose range should be chosen as wide as possible within the safety limit so that an adequate dose-response relationship can be adequately characterized. The number of dose levels, including the test drug product and the placebo, therefore, should be at least four. Ideally, it is preferable to select a dose level whose response is not expected to be statistically different from the placebo response so that the MED can be more accurately estimated. Sample size determination includes estimation of the total sample size and its distribution across different dose groups. Ruberg (1995a) suggested that the sample size be determined based on the statistical test for the hypothesis of interest for primary efficacy clinical endpoints. However, it should be noted that the sample size required for a dose-response study might be different from that for estimation of MED. Further research for the sample size for estimation of MED is needed.

In summary, a randomized, parallel dose-response trial is a widely used, successful, and acceptable design for obtaining population-average dose-response information. It should include three or more dosage levels, one of which may be placebo. For RPDRD, if dose levels are well chosen, the relationship of the test drug product with clinical efficacy or safety can be defined. However, it should be noted that the RPDRD with a placebo concurrent control would not be acceptable for the therapeutic areas such as life-threatening infections or potentially curable tumors. Because of the multiobjectives of RPDRD, statistical methods for estimation of the relationship between the dose and response such as

construction of confidence intervals and the use of graphical methods are equally important as the use of hypothesis testing procedures. The hypotheses for any RPDRD may be formulated as the natural ordering of the doses or as the particular questions regarding the shape of the dose-response curve with respect to different objectives of the study.

Crossover Dose-Response Design

Although RPDRD is a widely employed design for obtaining the population average dose-response relationship, it cannot provide the individual dose-response information. In addition, a critical understanding of different doses for a particular test drug product by any clinical researchers, biostatisticians, primary care physicians, or patients is that any given dose provides a mixture of desirable and undesirable effects, with no single dose necessarily optimal for all patients. Therefore, individual dose-response information is very important in real clinical practice for both physicians and patients. Therefore, if a drug effect develops rapidly and patients return to baseline conditions quickly after termination of therapy, and if responses are not irreversible, a randomized multiple-period crossover study of different doses using Williams' design (Chow and Liu, 2000) can be successful. Because for this design each individual patient receives several different doses of the test drug product, in addition to the population average dose-response curve, the distribution of the individual dose-response curves can be estimated.

Because the statistical inference for the population dose-response curve is based on the intrasubject variability, it may require fewer patients than the RPDRD. In addition, if the study is adequately designed and is well conducted with washout periods of sufficient length, dose and time will not be confounded together, and unbiased inference about the dose-response relationship can be obtained even in the presence of carryover effects. However, this design suffers the same shortcomings of all crossover designs. The duration of the study could be very long for each individual patient. Consequently, the likelihood of a patient's withdrawal during the studies is higher than RPDRD. In addition, there is often uncertainty about the carryover effects and about baseline comparability among different periods for potential period-by-treatment interaction. All of these in turn create analytic nightmares and questionable interpretation of the results.

Forced Titration (Dose-Escalation) Design

The forced titration design in ICH E4 guideline in fact employed the forced dose-escalation design described in Section 5.5, where all subjects move through the series of escalating doses. This design is similar to the crossover dose-response trials. However, assignment of subjects to receive different dose levels is ordered, not random. Therefore, a critical drawback of this design is that the dose effects of individual periods, time, and carryover effects (effects of cumulative doses) are all confounded together. No valid and unbiased statistical inference for the dose-response relationship based on continuous endpoints from this design is currently available. Forced titration trials may be the poor choice for collecting the dose-response information on adverse events because of their time-dependent nature. Other problems that make this design unattractive for the dose-response studies, including the poor choice for delayed response and spontaneous improvement over time. Consequently, the higher doses under this design may find very little room to show an incremental effect.

Similar to the crossover dose-response design, nonetheless, the forced titration design can provide a reasonable first approximation of both the population average dose-response curve and the distribution of individual dose-response relationships if the carryover effect is minimal and the number of dropouts is not excessive. In addition, if a placebo concurrent group is included in the forced titration trials, it allows a series of comparisons of an entire randomized group given several doses with a placebo concurrent control. In addition, because of inclusion of a placebo concurrent control, the forced titration design can provide evidence of effectiveness of the test drug product. Because the inference from this design may be based only on intrasubject variability, the forced titration design requires fewer patients than does the RPDRD. Because it can be used to investigate a wide range of doses, the forced titration design with a placebo concurrent group is a reasonable first dose-response study for helping selection of the doses for the subsequent randomized, parallel dose-response trials.

Optional Titration Design (Placebo-Controlled Titration to Endpoint)

The optional titration design in the ICH E4 guideline is the application of the titration design in Section 5.5 to investigate the dose-response relationship of the test drug product. The difference between the optional titration design and the forced-titration design, as described in Section 5.5, is that the subjects in the optional titration design are titrated up to the next higher doses if they fail to respond at the current dose level with respect to some prespecified efficacy criteria. Therefore, not all patients will receive all doses in the optional titration design. On the other hand, in the forced titration design, all patients will receive all doses if they do not experience any safety problems. Like the forced titration design, this design is applicable to indications where the response is quite prompt and is not an irreversible event, such as stroke or death. The design can provide information of population dose-response and individual dose-response curves. In addition, as shown in Table 5.5.2, by inclusion of a placebo concurrent group, the optional titration design not only can correct for spontaneous changes or investigator expectations, but also it can provide evidence of effectiveness of the test drug product.

Like the crossover design or forced titration design, this design suffers the same drawback that the time, dose, and carryover effects are confounded together in a complicated yet unknown pattern. Because only the poor responders are titrated up to receive higher doses, the responses observed at various dose levels are treatment-related, and hence, a crude analysis of the optional titration design will usually give a misleading inverted *U-shape* response-dose shape. This often leads to select a dose that is well in excess of what is really necessary, e.g., diuretics for hypertension or zidovudine (AZT) for acquired immune deficiency syndrome (AIDS). Currently, no valid statistical method for characterization of the dose-response relationship is available for the continuous response obtained from the optional titration design. However, this design mimics the real clinical practice and it requires much fewer patients than does the RPDRD. The optional titration design may be especially valuable as an early phase II study to identify doses for definitive parallel studies.

Many factors should be considered for selection of designs, dose levels, endpoints, and patient population in dose-response trials. These factors include the phase of clinical development, the therapeutic indications, and severity of disease in the target patient population. As pointed out by the ICH E4 guideline, the lack of appropriate salvage treatment of life-threatening or serious diseases with irreversible outcomes may preclude conduct of studies

below the maximum tolerable dose (MTD). If the target patient population is very homogeneous because of stringent inclusion/exclusion criteria, fewer patients may be required to achieve the objectives of the dose-response trials. However, the generalizability of the results from a homogeneous patient population is limited. On the other hand, larger and diverse patient populations allow detection of potentially important covariates and a much broader generalization of the results.

From a practical viewpoint, if the endpoints are continuous or categorical variables for some indications (e.g., blood pressure, analgesia, bronchodilation) that can be quickly obtained after the treatment is started and is rapidly dissipated after treatment is terminated, a wider range of designs can be used, and simple and small studies can provide useful dose-response information. Under these circumstances, the optional titration design with a placebo concurrent group is the typical design selected for many dose-response trials during early phase II clinical development to provide guidance for more definitive, randomized, parallel dose-response trials. In contrast, crossover or various titration designs are not applicable when the endpoints for some indications are delayed, persistent, or irreversible, e.g., stroke, asthma prophylaxis, survival in cancer, or treatment of depression. A randomized, parallel dose-response design is usually required for these indications. In addition, RPDRD offers protection against missing an effective dose because of an umbrella or bell-shape dose response curve, where higher doses are less efficacious than are lower doses (ICH E4, 1994).

It is prudent to conduct dose-response studies at the early stage of clinical development to reduce the number of failed phase III trials because of poor choice of the dose, or to avoid accumulation of a database that consists mainly of exposures at ineffective or excessive doses. It can also speed up the drug development process, and save development resources and cost. However, the endpoints of the dose-response studies may be different during different stages of clinical development. For example, as indicated by the ICH E4 guideline, in treatment of heart failure, a pharmacodynamic endpoint such as cardiac output, pulmonary capillary wedge pressure might be used in the early stage. An intermediate endpoint such as exercise tolerance and symptoms might be employed in the later development. However, a mortality or irreversible mortality endpoint such as survival or new infarction may be used for the final assessment.

7.6 COMBINATION TRIALS

As was indicated in Chapter 1, a treatment is defined as a combination therapy if it consists of more than one active ingredient. A treatment is called a combined product if it is a combination of different pharmaceutical entities such as drugs, biologics, and/or medical devices. Since the general principles of statistical designs and analyses for the assessment of the effectiveness and safety are the same for both combination therapy and combined product, in this section our discussion will be focused on the evaluation of a fixed dose for the combination of two or more active ingredients.

In recent years the search and development of combination therapy for various diseases have become a popular issue in the pharmaceutical industry (e.g., see Lasagna, 1975; Mezey, 1980). In clinical practice it is not uncommon to observe an enhanced therapeutic effect for a combination drug in which each component (at a certain dose) acts through a different mechanism when applied alone. For example, the combination of a diuretic with a beta-blocker has an enhanced therapeutic effect for the treatment of patients with hypertension. In practice,

Table 7.6.1 Maximum Tolerable Doses of Single Agents for Treatment of Advanced Ovarian Cancer

Drug	Maximum Tolerable Dose (mg/m ² /week)	Organ System of Dose-Limiting Toxicity
Cisplatin	35	Renal
Carboplatin	100	Bone marrow
Cyclophosphamide	400	Bladder
		Bone marrow
Hexamethylmelamine	2200	Gastrointestinal
Doxorubicin	30	Bone marrow
Paclitaxel	250	Neutropenia

Source: Adapted from Table 1 in Simon and Korn (1991) and Table 1 in Rowinsky and Donehower (1995).

however, each component of the combination drug must be administered alone at a high-dose level in order to achieve the desired therapeutic effect. In this case the component drugs may cause severe and/or sometimes fatal adverse events. Therefore it is desirable to control the synergistic effect among these components so that a combination at the lower-dose levels will reach the same or better effectiveness with less severe clinical toxicity. For example, Table 7.6.1 lists the conventional maximum tolerable doses for chemotherapeutic agents in the treatment of patients with advanced ovarian cancer. Each of these drugs induces serious organ-specific dose-limiting toxicity that jeopardizes the efficacy when they are administered alone. It is therefore imperative to search for different combinations of these agents at lower doses so that a higher equivalent cytotoxic dose can be achieved to provide a clinically meaningfully improvement in response rate and overall survival. For example, McGuire et al. (1996) report that paclitaxel and cisplatin provided not only a significantly higher response rate than a standard therapy of cisplatin plus alkylating agent cyclophosphamide (73% versus 60%, *p*-value = 0.01) for stage III and stage IV ovarian cancer but also a significant improvement in median survival (38 vs. 24 months, *p*-value < 0.001).

After the initial successful culture of *Helicobacter pylori* (or *H. pylori*) in gastric biopsy specimens from patients with histologic gastritis in 1982 (Warren, 1982), it was recognized that more than 95% of patients with duodenal ulcers and more than 80% of patients with gastric ulcers are infected with *H. pylori* (Walsh and Peterson, 1995). Recently the Consensus Development Conference on the U.S. National Institute of Health recommended that antimicrobial therapy be used to treat patients with ulcers who are also infected with *H. pylori* (NIH Consensus Development Panel, 1994). However, it is not easy to identify effective agents to eradicate the infection of *H. pylori*. The standard agents such as H₂-receptor antagonists and sucralfate have no effect on *H. pylori*. In addition, bimuth such as pepto-mismol and numerous antibiotics including erythromycin, amoxicillin, and metronidazole have not been proved to provide satisfactory long-term eradication rate (Peterson, 1991). As an alternative, a different regimen for the combination of antimicrobial agents and the combination of antimicrobial and antisecretory agents has been proposed for the treatment of patients with peptic ulcer disease who are infected with *H. pylori*. Table 7.6.2 displays various combinations of antimicrobial and antisecretory agents in the eradication of *H. pylori*. As can be seen from Table 7.6.2 these different combinations consist of different microbial and antisecretory agents at different dose levels.

Table 7.6.2 Combinations Used in the Eradication of *H. pylori*

Combinations	Dose	Duration	Eradication Rate (%)	Overall Adverse Events (%)
Bismuth	564 mg/4 times/day	14–15 days	89%	32%
Metronidazole	0.6–1.5 g/day	14–15 days		
Tetracycline	500 mg/4 times/day	14–15 days		
Bismuth	564 mg/4 times/day	10–14 days	84%	31%
Metronidazole	1.0–1.5 g/day	10–14 days		
Amoxicillin	1.5–2.0 g/day	10–14 days		
Metronidazole	500 mg 3 times/day	12 days	89%	13%
Amoxicillin	750 mg 3 times/day	12 days		
Ranitidine	300 mg at bedtime	6 weeks		
Clarithromycin	500 mg 3 times/day	10 days	86%	34%
Amoxicillin	750 mg 3 times/day	10 days		
Ranitidine	300 mg at bedtime	6 weeks		
Metronidazole	400 mg 3 times/day	14 days	90%	13%
Amoxicillin	750 mg 3 times/day	14 days		
Omeprazole	40 mg/day	6 weeks		
Metronidazole	500 mg 2 times/day	14 days	88%	18%
Clarithromycin	250 mg 2 times/day	14 days		
Omeprazole	20 mg 2 times/day	14 days		

Source: Table 1 in Walsh and Peterson (1995).

The duration for the treatment of these combinations and the corresponding eradication rates of *H. pylori* vary from one combination to another with different overall incidence rates of adverse events. In addition to the eradication rate of *H. pylori*, other important efficacy endpoints to be evaluated include the rate of complete healing of the ulcer and the recurrence rate as documented by endoscopy.

Note that although the FDA has a specific policy for the approval of combination therapy, the assessment of the effectiveness and safety for the combination therapy is rather difficult and yet challenging because of its complexity. The development of a combination therapy involves the determination of the optimal joint region of therapeutic dosing ranges, the duration of treatment for each different drug, and the order of administration of these agents, while the assessment of the combination therapy is usually based on multiple efficacy and safety clinical endpoints. This complexity may limit the chance of a combination therapy being approved by the FDA.

Fixed-Combination Prescription Drugs

For fixed-combination prescription drugs for humans, the FDA has specific regulations that are described in Section 300.50 in Part 21 of CFR. The regulation states that “Two or more drugs may be combined in a single dosage form when each component makes a contribution to the claimed effects and the dosage of each component (amount, frequency, duration) is such that the combination is safe and effective for a significant patient population.”

A fixed-combination prescription drug is defined as a single dosage formulated from different pharmacological agents. According to the FDA's regulation, the fixed-combination drugs are confined to a class in which different drugs can be formulated into one dosage form. However, different drugs may be administered in separate and/or different dosage forms either simultaneously or sequentially. For example, the Second International Study of Infarct Survival (ISIS-2, 1988) evaluated the effectiveness of the combined therapy of one-hour intravenous infusion of 1.5 MU of streptokinase up to 24 hours from the onset of chest pain followed by one month of oral enteric-coated aspirin at a dose of 150 mg per day for patients with suspected myocardial infarction. The dosage forms and routes of administration of the two therapeutic agents considered in ISIS-2 are different and are given sequentially. It should be noted that current regulations for fixed-combination drugs do not cover the combination therapy with different dosage forms and time of administration though, in principle, the statistical design and analysis are the same for both fixed-combination prescription drugs and combination therapy.

Note that for approval of a combination therapy, the FDA also requires the evidence for which each component must make a contribution to the claimed effects of the combination be provided. As indicated in Laska and Meisner (1989) and Hung et al. (1989, 1990, 1993), the *contribution* is referred to as the contribution based on a single clinical endpoint. On the other hand, for approval of a fixed-combination prescription drug, the FDA requires evidence to be provided of the superiority of the combination over each of its components. It, however, should be noted that the combination might increase the risk of adverse events due to the unnecessary exposure of patients to some components that may not have additional clinical benefit. In addition the FDA regulation states the relationship of the combination and its constituents in terms of the claimed effects. However, it does not specify the type, form, and magnitude of the contributions of each component to the claimed effects, nor does it require the claimed effects to be contrasted with a concurrent placebo control.

Let A and B be the components of a combination drug, denoted by $A + B$, each at a fixed dose. Let μ_A , μ_B , and μ_{AB} be the average of the distributions of the clinical endpoints such as reduction in seated diastolic blood pressure, eradication rate, vascular mortality rate within a certain time interval, or median survival. Since the FDA requires that the efficacy of the combination be superior to those of each component, the hypotheses can be formulated as follows (e.g., see Laska and Meisner, 1989; Hung et al., 1990):

$$\begin{aligned} H_0: \mu_A &\geq \mu_{AB} \quad \text{or} \quad \mu_B \geq \mu_{AB}; \\ \text{vs. } H_a: \mu_A &< \mu_{AB} \quad \text{and} \quad \mu_B < \mu_{AB}. \end{aligned} \tag{7.6.1}$$

Note that a larger value of μ indicates a better efficacy. Similar to the interval hypotheses described in (2.6.3), hypotheses (7.6.1) can be further decomposed into the following two one-sided hypotheses:

$$\begin{aligned} H_{01}: \mu_A &\geq \mu_{AB}; \\ \text{vs. } H_{a1}: \mu_A &< \mu_{AB}; \end{aligned} \tag{7.6.2}$$

and

$$\begin{aligned} H_{02}: \mu_B &\geq \mu_{AB}; \\ \text{vs. } H_{a2}: \mu_B &< \mu_{AB}. \end{aligned} \tag{7.6.3}$$

The first set of one-sided hypotheses is to verify that the combination drug $A + B$ is superior to component A . If the combination drug is indeed superior to component A , then component B does make a contribution of the claimed effect of the combination. Similarly the second set of one-sided hypotheses is to demonstrate that the combination $A + B$ is superior to component B . As a result, the null hypothesis (7.6.1) is a union of the two null hypotheses (7.6.2) and (7.6.3) which include all points in quadrants two, three, and four as well as all axes. On the other hand, the alternative hypothesis is the intersection of the two alternative hypotheses (7.6.2) and (7.6.3) which consists of all points in the first quadrant excluding the positive axes. Although hypotheses (7.6.1) can be employed to test statistically whether each component makes a contribution to the claimed effects of the combination, the type, form, and magnitude of the joint contribution by both components cannot be quantified. For example, consider the results of vascular mortality rates in days 0–35 reported by ISIS2, as shown in Table 2.4.2. The vascular mortality rates of IV placebo + active aspirin, IV streptokinase + aspirin placebo, and IV streptokinase and active aspirin are 10.7%, 10.4%, and 8.0%, respectively. The contribution of IV streptokinase to the combination is 2.7% ($10.7\% - 8.0\%$) and contribution of active aspirin to the combination is 2.4% ($10.4\% - 8.0\%$). However, without including a concurrent placebo control, it is not clear what the joint contribution of both drugs is. The mortality rate of IV placebo and aspirin placebo is 13.2%. The reduction in mortality rate of the combination therapy as compared to the concurrent placebo is 5.2% which is approximately the sum of the contribution of each component ($2.7\% + 2.4\%$). The estimated effect of combination versus placebo can be expressed as follows:

$$SA - PP = (SA - PA) + (SA - SP) + [(SP - PP) - (SA - PA)], \quad (7.6.4)$$

where S , A , and P stand for mortality rates of IV streptokinase, oral aspirin, and the placebo. The last term on the right-hand side of (7.6.4) is the interaction between the two components, which can be measured by the difference between the difference in vascular mortality between the streptokinase and placebo for patients taking the aspirin placebo tablets and that between the streptokinase and placebo for patients taking the active aspirin tablets. It turns out that the reduction in mortality of the combination as compared to the concurrent placebo is the sum of the contributions to the reduction in mortality by each component plus the interaction between two components as demonstrated by the mortality rates from ISIS2. That is,

$$\begin{aligned} 8.0\% - 13.2\% &= (8.0\% - 10.7\%) + (8.0\% - 10.4\%) \\ &\quad + [(10.4\% - 13.2\%) - (8.0\% - 10.7\%)] \\ &\quad - 5.2\% = -2.7\% - 2.4\% - 0.1\%. \end{aligned}$$

It can be seen from the above that the interaction (0.1%) between IV streptokinase and oral aspirin in the reduction of mortality is negligible. As a result, the magnitude of the contributions made by each component to the reduction in vascular mortality is quite similar, and their joint contribution to vascular mortality of the combination is additive. In practice, it is recommended that a complete 2×2 factorial design, as given in panel A of Table 7.6.3, be employed to assess the efficacy and safety of the combination with the two components each at a fixed dose. One could argue that there is no need for a concurrent placebo control in the evaluation of a fixed-combination drug because the clinical benefits of monotherapy for each component have been established through prospective, randomized, double-blind,

Table 7.6.3 Factorial Design for Combination Therapy

*Panel A: A Full 2×2 Factorial Design for Combination Therapy
of Two Components Each at Two Dose Levels*

Group	Drug A	Drug B
1	Placebo	Placebo
2	Placebo	Fixed active dose
3	Fixed active dose	Placebo
4	Fixed active dose	Fixed active dose

*Panel B: A Full $2 \times 2 \times 2$ Factorial Design for Combination Therapy
of Three Components Each at Two Dose Levels*

Group	Drug A	Drug B	Drug C
1	Placebo	Placebo	Placebo
2	Placebo	Placebo	Fixed active dose
3	Placebo	Fixed active dose	Placebo
4	Placebo	Fixed active dose	Fixed active dose
5	Fixed active dose	Placebo	Placebo
6	Fixed active dose	Placebo	Fixed active dose
7	Fixed active dose	Fixed active dose	Placebo
8	Fixed active dose	Fixed active dose	Fixed active dose

and placebo-controlled studies. This information, however, can only be served as a historical control for external validation for the assessment of a combination drug. Without a concurrent placebo control, the clinical benefit of the combination under evaluation may never be internally validated. In addition the dose levels of the combination assessed by clinical trials are usually smaller than those selected for monotherapy in order to produce synergistic effects that achieve better efficacy and safety. Therefore effectiveness of the dose levels of each component may not be adequately evaluated alone before as a single agent. As a result a concurrent placebo control must be used to investigate and quantify the effect of each component and the type, form, and magnitude of their contributions to the claimed effects of the combination under study.

Hypotheses (7.6.1), which can be tested based on the data from phase III studies, are often used to confirm the effectiveness and safety of the combination drug for each component at a particular dose level. The doses of each component in a combination drug are often determined at the stage of phase II clinical development. The combination can then be confirmed through phase III clinical trials. In practice, it usually requires sophisticated and state-of-art statistical design and analysis for identification of possible combinations of doses at which the contribution of each component to the claimed effects can be assessed.

Multilevel Factorial Design

A commonly employed design for the assessment of combination drugs or therapy is the factorial design with multilevels, which is known as a multilevel factorial design. For assessment of a combination drug with a factorial design, each component of the combination drug is referred to as a *factor*, and the doses of a component are the *levels* of the corresponding factor. Therefore a full multilevel factorial design includes a number of

Table 7.6.4 A Full $(a + 1) \times (b + 1)$ Factorial Design for Combination Therapy of Two Components at a and b Dose Levels

Group	Drug A	Drug B
1	Placebo	Placebo
2	Placebo	Active dose 1
:	:	:
$b + 1$	Placebo	Active dose b
$b + 2$	Active dose 1	Placebo
$b + 3$	Active dose 1	Active dose 1
:	:	:
$2(b + 1)$	Active dose 1	Active dose b
$a(b + 1)$	Active dose a	Placebo
$a(b + 2)$	Active dose a	Active dose 1
:	:	:
$(a + 1)(b + 1)$	Active dose a	Active dose b

treatment groups which is made up of all possible combinations of factors and levels. The number of treatment groups equals the product of the number of levels and the number of factors. For example, in panel A of Table 7.6.3 there are two component drugs, and each component drug has two dose levels including dose 0 (or placebo) and one active fixed dose level. As a result this factorial design has two factors and each has two levels, so there are a total of four treatment groups. A design of this kind is denoted by a 2×2 factorial design. As another example, panel B of Table 7.6.3 displays a $2 \times 2 \times 2$ factorial design for a combination of three component drugs where each component has two dose levels (i.e., dose 0 or a placebo and a selected active dose level). Therefore there are a total of eight treatment groups. Note that the statistical designs employed for panels A and B in Table 7.6.3 are parallel-group designs in which the number of parallel arms equals to the number of treatment groups. Therefore, when a factorial design is employed, the number of parallel groups can be as large as the number of factors and/or the number of levels of each factor increase. For example, Table 7.6.4 provides the treatment groups of clinical trials that assess a combination drug that consists of component drugs *A* and *B*. If we include dose 0 (or placebo) and a active dose levels for *A* and b active dose levels for *B*, then there are a total of $(a + 1)(b + 1)$ treatment groups. Hence the number of treatment groups would be 4, 9, 16, or 25 when the number of active dose levels for each component are 1, 2, 3, and 4, respectively. In addition a parallel-group factorial design usually requires a large sample size. As an alternative, some crossover designs such as Williams design or balanced incomplete block design might be useful. Table 7.6.5 provides the Williams design for a 2×2 factorial design with and without a concurrent placebo control group.

For a combination drug consisting of component drugs *A* and *B*, in the case where the effectiveness and safety of component *A* has been established for the intended indication, then the selection of the dose of component *B* is critical in providing an enhanced therapeutic effect. In practice, it is suggested that the dose-response relationship of component *B* at an established effective dose of component *A* be examined. In some cases, since patients who receive component *A* may exhibit some differential limiting clinical toxicity, it is suggested that doses of component *B* be administered at a preselected dose after the patients are titrated to the maximum tolerable dose of component *A*. A typical example is

Table 7.6.5 Williams Designs for Combination Therapy

Sequence	Period		
	I	II	III
1	A	B	A + B
2	B	A + B	A
3	A + B	A	B
4	A	A + B	B
5	B	A	A + B
6	A + B	B	B

Panel A: 2 × 2 Factorial With Concurrent Placebo Control				
Sequence	Period			
	I	II	III	IV
1	Placebo	A + B	A	B
2	A	Placebo	B	A + B
3	B	A	A + B	Placebo
4	A + B	B	Placebo	A

the study of the combination therapy of cyclosporine and methotrexate for patients with severe rheumatoid arthritis (Tugwell et al., 1995). Note that the type of factorial design employed in this study is usually referred to as a partial factorial design because it does not utilize all possible combinations of dose levels from each factor as indicated in Table 7.6.6.

One important advantage of a combination therapy is that it is usually able to achieve a better clinically significantly synergistic efficacy with fewer adverse events at lower-dose levels compared to individual monotherapy of its components. In practice, although each

Table 7.6.6 Designs for Two-Drug Combined Therapy With a Fixed Dose of One Component

Panel A: The Same Fixed Dose for One Component		
Group	Drug A	Drug B
1	Active dose 1	Placebo
2	Active dose 1	Active dose 1
:	:	:
b + 1	Active dose 1	Active dose b

Panel B: Titration of The Dose Level of One Component to The Maximum Tolerable Dose		
Group	Drug A	Drug B
1	Titrated to MTD	Placebo
2	Titrated to MTD	Active dose 1
:	:	:
b + 1	Titrated to MTD	Active dose b

monotherapy of the component drugs may also achieve the same efficacy at a higher dose, the corresponding safety may not be tolerable.

Global Superiority of Combination Drug

In order to identify the optimal combination of dose levels and to investigate the potential drug-to-drug interactions, it is recommended that several or all possible combinations be explored simultaneously with multilevel factorial designs. Table 7.6.7 provides a cross tabulation of a full multilevel factorial design with μ_{ij} representing the average response of the combination made of dose i of component A and dose j of component B. A separate dose-response relationship of each component can be investigated by the first row for drug B or first column for drug A where the placebo is administered for the other component. In order to provide a useful dose-response relationship for therapeutic applications, Hung et al. (1989) suggests that the dose ranges to be investigated must include a very low dose level and a very high dose level that are not in the effective dose range. As a result, the contribution of some doses by one component when added to the other drug may not be different from the placebo, while the contribution of the other component is quite obvious. Hence with respect to the requirement described in Section 300.50 in Part 21 of CFR that each component makes a contribution to the claimed effects, Hung et al. (1990) defines the superiority of a combination drug over its component drugs in a global sense. Strict superiority of a combination drug is defined as the existence of at least one dose combination that is more effective than its components. Let d_{ij} be the minimum gain in efficacy obtained from combining dose i of drug A and dose j of drug B as compared to its component drugs at the same dose levels alone,

$$d_{ij} = \mu_{ij} - \max(\mu_{i0}, \mu_{0j}), \quad i = 1, \dots, a, j = 1, \dots, b. \quad (7.6.5)$$

The corresponding statistical hypotheses can then be formulated as:

$$\begin{aligned} H_0: d_{ij} &\leq 0 \quad \text{for every } 1 \leq i \leq a \quad \text{and} \quad 1 \leq j \leq b; \\ \text{vs. } H_a: d_{ij} &> 0 \quad \text{for some } 1 \leq i \leq a \quad \text{and} \quad 1 \leq j \leq b. \end{aligned} \quad (7.6.6)$$

A combination drug is said to be superior to its components in the wide sense if the average of all the dose combinations is superior to both the averages of the individual

Table 7.6.7 Cross Tabulation of a Full $(a + 1) \times (b + 1)$ Factorial Design for Combination Therapy with the Mean Effects

Dose Levels for Component A	Placebo	Dose Levels for Component B					Overall
		1	2	...	b		
Placebo	μ_{00}	μ_{01}	μ_{02}	...	μ_{0b}	μ_0	
1	μ_{10}	μ_{11}	μ_{12}	...	μ_{1b}	μ_1	
2	μ_{20}	μ_{21}	μ_{22}	...	μ_{2b}	μ_2	
:	:	:	:	...	:	:	
a	μ_{a0}	μ_{a1}	μ_{a2}	...	μ_{ab}	μ_a	
Overall	μ_0	μ_1	μ_2	...	μ_b	$\mu..$	

monotherapy doses. Let m_A represent difference between the average responses of all (i, j) combinations over the entire range of active doses and that dose i of component A when placebo is administered for component B ,

$$m_A = \text{Ave}(\mu_{ij} - \mu_{io}).$$

Similarly

$$m_B = \text{Ave}(\mu_{ij} - \mu_{oj}).$$

Consequently the concept of the wide global superiority of combination drug can be stated as:

$$\begin{aligned} H_0: m_A \leq 0 &\quad \text{or} \quad m_B \leq 0; \\ \text{vs.} \quad H_a: m_A > 0 &\quad \text{and} \quad m_B > 0. \end{aligned} \quad (7.6.7)$$

The strict global superiority of a combination drug restricts our attention to a more effective class of combinations than either component drug alone. According to d_{ij} in the alternative hypothesis of (7.6.6), we only need to identify one μ_{ij} that is better than mean responses μ_{io} and μ_{oj} of the corresponding monotherapy of component A at dose i and of component B at dose j . However, it does not guarantees superiority of one combination over all dose levels of either monotherapy.

Consider the example given in Hung et al. (1993). Table 7.6.8 displays the mean reductions from the baseline in post-treatment supine diastolic blood pressure from a clinical trial that evaluated a combination drug with three dose levels for drug A and four dose levels for drug B , including the placebo (Hung et al., 1993). From Table 7.6.8 the mean reductions in diastolic blood pressure of all combinations are seen to be more than those at the corresponding doses of either monotherapy. As a result, the minimum gain d_{ij} are positive for all combinations, which is shown in Table 7.6.9. From Table 7.6.10 suppose that the mean reduction from the baseline in diastolic blood pressure of dose 3 for component B as monotherapy is changed from 3 to 7. Table 7.6.11 reveals that although the minimum gain of all combinations over their corresponding monotherapy is at least 0, only combinations (1, 1) and (1, 3) surpass the monotherapy of drug B at dose level 3. This result is desirable because the monotherapy of drug B at the highest dose may induce some severe adverse events, while the dose level of drug B in combination with a larger reduction is two levels

Table 7.6.8 Difference From Placebo in Mean Reduction in Supine Diastolic Blood Pressure (mmHg)

Dose Levels for Component A	Dose Levels for Component B			
	Placebo	1	2	3
Placebo	0	4	4	3
1	5	9	7	8
2	5	6	6	7

Source: Hung et al. (1993).

Table 7.6.9 Minimum Gain of Combinations for the Data in Table 6.4.8

Active Dose Levels for Component A	Active Dose Levels for Component B		
	1	2	3
1	4	2	3
2	1	1	2

Source: Hung et al. (1993).

lower than that of the monotherapy. As a result the combination of dose 1 of drug A with dose 1 of drug B provides a much safer margin and hence a larger benefit-to-risk ratio.

For assessment of wide global superiority, as outlined in hypotheses (7.6.7), there seem to be, on the one hand, a combination that is better than its component A and, on the other, a combination that might not be the same combination but is better than its component B. Unlike strict superiority, however, this does not guarantee that there is a combination that is better than both of its components as required by the FDA regulation. Therefore the strict global superiority meets the current regulatory requirement for approval of a combination drug product. Hung et al. (1990, 1993) propose two statistical testing procedures for hypotheses (7.6.6) under the assumption that data are normally distributed. The proposed methods reduce to the min test for hypotheses (7.6.1) proposed by Laska and Meisner (1989) when only one active dose level is included in the assessment trial of a combination drug. Note that the sampling distributions of the methods proposed by Hung et al. are quite complicated and require special tables for the significance tests. Hung (1996) extends the application of these two methods to the situations (1) where the variance of the clinical endpoint is a function of its mean and (2) where an incomplete factorial design is used.

When a combination drug consists of more than two component drugs, the concept of strict global superiority for the combination drug of two components can be easily extended. Let d_{ijk} be the minimum gain in efficacy obtained by combining dose i of drug A with dose j of drug B, and dose k of drug C over its components alone at the same dose levels,

$$d_{ijk} = \mu_{ijk} - \max(\mu_{ioo}, \mu_{ijo}, \mu_{ook}), \quad i = 1, \dots, a, j = 1, \dots, b, k = 1, \dots, c, \quad (7.6.8)$$

where μ_{ijk} is the mean response of the combination of dose i of drug A, dose j of drug B, and dose k of drug C, and μ_{ioo} , μ_{ijo} and μ_{ook} represent the mean responses of its components A, B, and C alone at the same dose levels i , j , and k , respectively, where the active agents are administered with the placebos of the other two components. The strict global superiority

Table 7.6.10 Modified Differences From Placebo in Mean Reduction in Supine Diastolic Blood Pressure (mmHg)

Dose Levels for Component A	Dose Levels for Component B		
	Placebo	1	2
Placebo	0	4	4
1	5	9	7
2	5	6	6

**Table 7.6.11 Minimum Gain of Combinations
for the Modified Data in Table 7.6.10**

Active Dose Levels for Component A	Active Dose Levels for Component B		
	1	2	3
1	4	2	1
2	1	1	0

of a three-drug combination over its components can be formulated by the following statistical hypotheses:

$$\begin{aligned} H_0: d_{ijk} \leq 0 & \text{ for every } 1 \leq i \leq a, 1 \leq j \leq b, \text{ and } 1 \leq k \leq c; \\ \text{vs. } H_a: d_{ijk} > 0 & \text{ for some } 1 \leq i \leq a, 1 \leq j \leq b, \text{ and } 1 \leq k \leq c. \end{aligned} \quad (7.6.9)$$

The above hypotheses can be tested to verify the existence of strict global superiority of a combination drug. However, the extension of the methods of Hung et al. (1993) for testing the above hypotheses is not straightforward. Further research is needed.

Suppose that a combination of two component drugs is developed to treat patients with benign prostatic hyperplasia (BPH). Two primary efficacy endpoints for assessment of the combination are peak urinary flow rate (mL/s) and AUA-7 symptom scores. Let c_{ij} and d_{ij} be the minimum gain of combination of dose i of drug A with dose j of drug B of peak urinary flow rate and AUA-7 symptom scores, respectively. Following the suggestion by Laska and Meisner (1990), the strict global superiority for more than one clinical endpoint is defined as that where at least one combination is superior to its component drugs for at least one clinical endpoint. Hence, if a combination is better than its component drugs for at least one clinical endpoint, then the minimum gain of the combination based on both clinical endpoints must be greater than zero. The hypotheses corresponding to the strict global superiority can be formulated as follows:

$$\begin{aligned} H_0: \min(c_{ij}, d_{ij}) \leq 0 & \text{ for every } 1 \leq i \leq a \text{ and } 1 \leq j \leq b; \\ \text{vs. } H_a: \min(c_{ij}, d_{ij}) > 0 & \text{ for some } 1 \leq i \leq a \text{ and } 1 \leq j \leq b. \end{aligned} \quad (7.6.10)$$

The concept of hypotheses (7.6.10) can easily be extended to an evaluation of the combination drug based on more than two endpoints. However, the definition of strict global superiority and its corresponding formulation of hypotheses and proposed statistical procedures are to verify the existence of at least one combination that is better than both of its components. Furthermore they are hypothesis testing procedures and hence cannot describe the dose-response relationship and potential drug-to-drug interaction among components. Consequently they fail to provide a way to search for combinations for which each component makes a contribution to the claimed effects should they exist.

Method of Response Surface

To overcome the drawbacks associated with the definition of the strict global superiority for combination drug, the concept of response surface methodology can provide a nice compliment to the statistical testing procedures suggested by Laska and Meisner (1989)

and Hung et al. (1993). For a combination trial conducted with a factorial design, if the dose levels of each component are appropriately selected, then the technique of response surface can provide valuable information regarding (1) the therapeutic dose range of the combination drug with respect to effectiveness and safety and (2) the titration process and drug-to-drug interaction. The response surface method can empirically verify a model that adequately describes the observed data. For example, we can consider the following statistical model to describe the response:

$$Y_{ijk} = f(A_i, B_j, \theta) + e_{ijk}, \quad (7.6.11)$$

where Y_{ijk} is the clinical response of patient k who receives dose i of drug A , denoted by A_i and dose j of dose B , denoted by B_j , θ is a vector of unknown parameters, and e_{ijk} is the random error in observing Y_{ijk} , where $k = 1, \dots, n_{ij}, j = 1, \dots, b, i = 1, \dots, a$. The component $f(A_i, B_j, \theta)$ in (7.6.11) gives a mathematical description and an approximation to the true unknown response surface provided by the two drugs. When the primary clinical endpoints of interest are continuous variables, $f(A_i, B_j, \theta)$ is usually approximated by a polynomial. For a detailed description of response surface methodology, see Box and Draper (1987); also see Peace (1990) for applications of response surface to a phase II development of antianginal drugs.

If the assumed mathematical model is not too complicated, then the standard statistical estimation procedures such as least squares method or maximum likelihood method can be selected to estimate the unknown parameter θ . After substitution of the parameter in $f(A_i, B_j, \theta)$ by its estimate, an estimated response surface can be obtained that provides an empirical description of the dose-response relationship either by a three-dimensional surface or by two-dimensional contours. In addition an optimal dose combination can be estimated to give an maximum clinical response if it exists and is unique. If the 95% confidence region for the optimal dose combination does not lie on both the horizontal axis representing drug A and the vertical axis representing drug B at the contours, then it is concluded that the estimated optimal combination is superior to both of its components. The technique of response surface therefore can estimate an optimal dose combination that may not be the combination of doses of both components selected for the trial given existence of such combinations. Hung (1992) suggests a procedure to identify a positive dose-response surface for combination drugs. Hence the response surface can also estimate a region in which the combinations are superior to their components. Estimation of a superior region is particularly appealing if safety is the major reason for the combination drug because a combination can be chosen from the superior region with much lower doses of both components to achieve the same superiority in efficacy but with a much better safety profile. Since μ_{ij} is estimated using the information of the entire sample, if the assumed model can adequately describe the dose-response relationship of the combination, the response surface method requires fewer patients than the procedures for testing strict global superiority proposed by Hung et al. (1993). Note that estimation of the unknown parameters, response surface, and the optimal combination dose and construction of the confidence region for the optimal combination dose are model dependent. Consequently, as indicated by Hung et al. (1990), the FDA is concerned about the application of response surface method due to the following reasons:

1. The sensitivity of the methods such as lack-of-fit tests, goodness-of-fit tests, and residual plots for verification of the adequacy of the fitted models is often questionable.

2. Even if there is no evidence for the inadequacy of the fitted model, the chance of selecting an inadequate model and the effect of such an error cannot be evaluated.
3. The response surface method is model dependent. Two different models that both adequately fit the given data may provide contradictory conclusions in any subsequent statistical analyses.

The methods for strict global superiority and response surface methodology should play crucial but complimentary roles in assessment of combination drugs. Existence of a combination superior to its components can be verified first by the model-independent statistical testing procedures proposed by Hung et al. (1993). Then the response surface technique can be applied to (1) empirically describe the dose-response relationship, (2) identify for the region of superior efficacy, and (3) estimate the optimal dose combination. Also it would be of interest to provide a scientific justification as to why the two methods yield inconsistent conclusions.

7.7 BRIDGING STUDIES

7.7.1 Introduction

As indicated earlier, for marketing approval of a study medicine, the FDA requires that substantial evidence of the effectiveness and safety of the study medicine be provided through the conduct of at least two adequate and well-controlled clinical trials (the so-called pivotal trials). The purpose of at least two clinical trials is to confirm the reproducibility of the evidence on efficacy, safety, and dose response of the study medicine. After the study medicine is approved by the regulatory agency, the sponsor may seek registration of the approved medicine in a new region (e.g., European Community or Asian-Pacific countries). However, the possible differences in ethnicity, culture, and clinical practice between the original region where the study medicine is approved and the new region and their impacts on the safety, efficacy, dose, and dosing regimen have limited the willingness of the regulatory authority in the new region to accept the clinical data generated in the original region. As a result, the regulatory authority in the new region often requests the sponsor to repeat similar studies for obtaining all or much of the clinical data in the new region to confirm the reproducibility of the evidence on efficacy, safety, and dose response of the study medicine before the study medicine can be approved in the new region. This extensive duplication of clinical evaluation in the new region not only demands valuable development resources, but also it delays availability of the new medicine to needed patients in the new region. To resolve this dilemma, the ICH has recently published a tripartite guideline entitled, *Ethnic Factors in the Acceptability of Foreign Clinical Data*, which is usually referred to as the ICH E5 guideline, to address the above issues (ICH, 1998).

As stated in the ICH E5 guideline, the objective of the guideline is to provide a framework for evaluation of the impact of ethnic factors on the efficacy and safety of a study medicine at a particular dosage or dose regimen. In addition, it describes regulatory strategies of minimizing duplication for clinical data and requirement of bridging evidence for extrapolation of foreign clinical data to a new region. In this section, we will provide an overview of the ICH E5 guideline, including ethnic sensitivity, necessity of bridging studies, types of bridging studies, and the assessment of similarity between regions based on

bridging evidence. Also included is a brief discussion regarding challenges on the establishment of regulatory requirements, the assessment of bridging evidence, and the design and analysis of bridging studies.

7.7.2 Ethnic Sensitivity and Necessity of Bridging Studies

Ethnic Sensitivity The ICH E5 guideline lists critical properties of a compound for assessment of sensitivity to ethnic factors. These critical properties include linear pharmacokinetics (PK), flat pharmacodynamic (PD), therapeutic range, degree of metabolism, extent of bioavailability, potential for protein binding, potential for interactions, genetic polymorphism, intersubject variability, systemic mode of action, and potential for inappropriate use. However, the ICH E5 guideline also points out that no one property of the medicine is predictive of the compound's relative sensitivity to ethnic factors. Because of the complexity due to possible interaction among the drug's pharmacological class, indication, and demographic of patient population, the ICH E5 does not provide a precise and definitive criterion for evaluation of ethnic sensitivity. As a result, no probability statements can be made for the errors resulting from the decision making on sensitivity to ethnic factors. Therefore, both regulatory authority in the new region and the sponsor do not have a criteria and a method for an objective and impartial evaluation of ethnic sensitivity and the necessity of a bridging study.

Necessity of Bridging Studies As no well-defined and scientifically justifiable criteria for assessment of ethnic sensitivity are indicated in the ICH E5 guideline, any proposed approach for the assessment of the necessity of bridging studies could be subjective and controversial and may not be accepted by the regulatory authority and sponsors in the new region. However, several approaches have been proposed in the literature based on some statistical criteria. See, for example, Shao and Chow (2002), Chow et al. (2002), and Liu and Chow (2002). These approaches are briefly outlined below.

Shao and Chow (2002) proposed the use of the concept of reproducibility probability as a *statistical guide* for quantifying the likelihood that under the same experimental conditions, a second trial conducted in the new region can reproduce the same results of the first trial in the original region. The information of reproducibility probability provides a guide regarding statistical inference on the chance that clinical results observed in the original region are reproducible at the new region. Note that the reproducibility of a clinical trial is referred to as the probability that clinical results observed from a region are reproducible in the *same* targeted patient population in the same region or from region to region.

To address the possible impact of ethnic factors, Chow et al. (2002) suggested performing a sensitivity analysis on the reproducibility probability with respect to a sensitivity index Δ , which is defined as

$$\Delta = (1 + \varepsilon/\mu_d)/C,$$

where μ_d is the mean difference between the treatment groups observed in the original region, ε is the change of μ_d in the new region, and C represents the change in variability in the new region. Chow et al. (2002) refer to the reproducibility probability with respect to the sensitivity index as the generalizability probability. Note that the generalizability of a clinical trial is referred to as the probability that clinical results observed from a region are reproducible in a *similar but slightly different* (it might be due to a difference in ethnicity

or other factors such as age or sex) targeted patient population in the same region or from region to region. The concept of generalizability probability may be useful in providing regulatory authorities an opportunity to choose a statistical criterion for determining whether a clinical bridging study is necessary.

For an objective evaluation of the necessity of bridging studies, Liu and Chow (2002) suggested performing a meta-analysis for a systematic overview of ethnic sensitivities based on clinical data accumulated regarding the approved medicines. As indicated in the United States Drug Master File, thousands of medicines have been approved for various indications for different patient populations by various regulatory authorities in different geographic regions. Sufficient preapproval and postmarketing experience on the critical properties regarding ethnic sensitivities and the impact of intrinsic and extrinsic factors on efficacy, safety, dosage, and dose regimen have been accumulated for these medicines. Based on these data, an instrument consisting of three domains can be developed to determine the degree of the impact of difference in ethnic factors on the efficacy, safety, dose, and dose regimen of these medicines and consequently the necessity of bridging studies.

For example, the first domain may include the critical properties of the compound mentioned above. The second domain may consist of intrinsic factors as discussed in Appendix A of the ICH E5 guideline (see also Table 7.7.1). As described in the ICH E5 guideline, intrinsic ethnic factors are factors that help to define and identify a subpopulation and may influence the ability to extrapolate clinical data between regions. Examples of intrinsic factors include genetic polymorphism, age, gender, height, weight, lean body mass, body composition, and organ dysfunction. The third domain may comprise the extrinsic ethnic factors as described in Appendix A of the ICH E5 guideline (see also Table 7.7.1). As indicated in the ICH E5 guideline, extrinsic ethnic factors are factors associated with the

Table 7.7.1 Intrinsic and Extrinsic Ethnic Factors

Classification	Factor	Description
Intrinsic	Genetic	Gender Race Genetic polymorphism of the drug metabolism Genetic diseases
	Physiological and pathological conditions	Age (children-elderly) Liver Kidney Cardiovascular functions Disease
	Environmental	Climate Sunlight Pollution
	Culture	Socioeconomic factors Educational status Language
	Medical practice	Disease definition/diagnostic Therapeutic approach Drug compliance
	Regulatory practice	GCP Methodology/endpoints

environment and culture in which a person resides. Extrinsic factors tend to be less genetically and more culturally and behaviorally determined. Examples include the social and cultural aspects of a region such as medical practice, diet, use of tobacco, use of alcohol, exposure to pollution and sunshine, socioeconomic status, compliance with prescribed medications, and particularly important to the reliance on studies from a different region, practices in clinical trial design and conduct. Within each domain, a scoring scheme for each property or factors may be designed to characterize the degree of the impact on efficacy, safety, dose, and dose regimen. For example, a possible scoring scheme could be a 5-point system such as 1 (no), 2 (mild), 3 (moderate), 4 (strong), and 5 (complete).

An algorithm can then be developed to provide a summary index for an overall assessment of the impact on the efficacy, safety, dosage, and dose regimen of the study medicine. In practice, the database of these compounds can be divided into two date sets, namely, a training set and a validation set. Based on the summary indices computed from the medicine in the training set, a threshold can be determined to classify these medicines into two groups. One group consists of medicines that are insensitive to ethnic factors and hence do not require bridging studies. The other group contains medicines that are ethnic sensitive and hence require bridging studies. The probability of classification error can then be evaluated based on the validation set. Within the group of medicines with the necessity for bridging studies, further cutoff points can be estimated for different types of bridging studies. When a new medicine is applied for registration in the new region, the regulatory authority in the new region and the sponsor can calculate the summary index of the study medicine to determine whether a bridging study is necessary and what type of the bridging study is warranted. The above approach seems logical and acceptable to the pharmaceutical industry and regulatory agencies. However, it requires a joint effort and close collaboration among researchers/scientists from multiple disciplines in pharmaceutical research and development.

7.7.3 Types of Bridging Studies

In general, the types of bridging studies required depend on the ethnic sensitivity of the study medicine, clinical experience of the drug class, extrinsic ethnic factors, and ethnic differences between the new and original regions. As stated in the ICH E5 guideline, a bridging study could be a pharmacokinetic/pharmacodynamic (PK/PD) study or a controlled clinical trial (CCT). Table 7.7.2 provides a summary of types of bridging studies with respect to the factors of region, medical practice, drug class, and clinical experience.

As can be seen from Table 7.7.2, there are at least two fundamental issues in the ICH E5 guideline: (1) sensitivity of study medicines to ethnic factors, necessity of a bridging study,

Table 7.7.2 Types of Bridging Studies

Medicine	Region	Medical Practice	Drug Class	Clinical Experience	Bridging Studies
Insensitive	—	Similar	—	—	No
Sensitive	Similar	—	—	Sufficient	No
Sensitive	Dissimilar	Similar	Familiar	—	PD
Choice of Dose	—	Different	Unfamiliar	Insufficient	CCT

PD = Pharmacodynamics

CCT = Controlled clinical trials

and the nature and type of bridging studies; and (2) assessment of similarity based on bridging evidence.

7.7.4 Assessment of Similarity Based on Bridging Evidence

According to the ICH E5 guideline, a bridging data package consists of (1) selected information from the so-called complete clinical data package (CCDP) that is relevant to the population of the new region, and (2) if needed, a bridging study to extrapolate the foreign efficacy and/or safety data to the new region. In other words, bridging evidence is actually provided either in the CCPD generated during clinical drug development program for submission to the original region or in a bridging study conducted in the new region after the pharmaceutical product is approved in the original region. When the bridging evidence provided in the CCPD could not allow extrapolation of foreign clinical data to a new region, then a bridging study should be conducted in the new region to generate a limited amount of clinical data to bridge the clinical data between the two regions. The ICH E5 guideline clearly states that assessment of the ability of extrapolation of the foreign data rely on the *similarity* of dose response, efficacy, and safety between the new and original regions, either with or without dose adjustment. However, the ICH E5 guideline does not provide a precise definition or criteria for evaluation of similarity. For assessment of similarity, a number of different statistical procedures have been proposed based on different definitions or concepts of similarity, for example, batch similarity in stability analysis for shelf-life estimation, similarity in drug release for comparison of dissolution profiles between drug products, similarity in drug absorption for assessment of bioequivalence between drug products, and the concept of consistency between clinical results. Among these definitions and concepts of similarity, bioequivalence in drug absorption and consistency between clinical results are most relevant.

Population Similarity Based on the concept of similarity in drug absorption for assessment of bioequivalence between drug products, Chow et al. (2002) proposed to adopt the concept of population bioequivalence for assessment of similarity of clinical results between the original region and the new region. The idea is to establish similarity according to the following aggregated equivalence (similarity) criteria, which is similar to the one proposed in the recent FDA guidance for establishment of population and individual bioequivalence (FDA, 2001a):

$$\theta = \frac{(\mu_0 - \mu_1)^2 + \sigma_{T1}^2 - \sigma_{T0}^2}{\sigma_{T0}^2},$$

where $\sigma_{Tk}^2 = \sigma_{Bk}^2 + \sigma_{Wk}^2$ is the total variance in region k , $k = 0,1$ and σ_{Bk}^2 and σ_{Wk}^2 are the between-center variance and the within-center variance in region k . We claim that clinical results in the new region are similar to those observed in the original region if the 95% upper confidence bound of the following linearized criterion is less than 0:

$$\varsigma = (\mu_0 - \mu_1)^2 + \sigma_{T1}^2 - (1 + \theta_U)\sigma_{T0}^2,$$

where θ_U is the similarity margin set by the regulatory authority in the new region.

Consistency Among Studies For the concept of consistency between clinical studies, Shih (2001) suggested the use of a predictive probability function for measuring the consistency

between clinical results from different studies. His approach is briefly outlined below. Suppose that there are a total of H reference studies and each compares the same two treatment groups. Let \mathbf{W} denote the vector of standardized between-group differences of these H reference studies $\mathbf{W} = (w_1, \dots, w_H)$ and v is the standardized between-group difference for the bridging study. The result of v from the bridging study is consistent with the previous results \mathbf{W} if and only if

$$p(v|\mathbf{W}) \geq \min\{p(w_h|\mathbf{W}) | h = 1, \dots, H\},$$

where $p(a|\mathbf{W})$ is the predictive probability function that provides a measure of plausibility, given the previous results \mathbf{W} .

Hierarchical Model Under a hierarchical model, on the other hand, Liu et al. (2002) proposed the use of the concept of equivalence or noninferiority for evaluation of similarity between the new and original regions. Their approach is summarized below:

- Step 1: From the complete clinical data package, under a hierarchical model, use the clinical data from the original region to obtain the estimate of relative efficacy and its estimated standard error.
- Step 2: From the data of the bridging study, obtain the estimate of relative efficacy and its estimated standard error in the new region.
- Step 3: Based on the estimated relative efficacy and its standard error from both new and original regions and equivalence limit, perform the usual two one-sided tests procedure or one-sided noninferiority procedure (or confidence interval).

However, one can argue that the results of clinical data in the new region are similar to those in the original region if the new region also demonstrates a positive treatment effect.

Bayesian Approach Liu et al. (2002) suggested an empirical Bayesian approach to provide the evidence of a positive treatment effect. The results on dose response, efficacy, or safety of the original region can be incorporated as a prior to evaluate a positive treatment effect by the bridging data in the new region. In other words, for a prespecified significance level α , $0 < \alpha < 1$, one can conclude similarity based on the concept of a positive treatment effect if the posterior probability

$$P\{\text{a positive treatment effect} | \text{data and prior}\} > 1 - \alpha.$$

Despite the above different definitions for similarity, a direct interpretation of the ICH E5 guideline on similarity requires performing a between-region (study) analysis to evaluate the treatment-by-region interaction. It is then very clear that the sample size required for the test based on the treatment-by-region interaction will be much larger than that for detection of the treatment effect alone (Liu et al. 2002). This statement is true for all types of studies and for all types of endpoints. On the other hand, one only wants to verify whether the evidence of efficacy or safety or PK/PD properties observed in the original region can be reproduced in the new region. In this context, for example, a statistical significance based on a particular endpoint can be also obtained from the bridging study conducted in the new region if it had been observed in the original region. However, an equal or even larger sample size is required to reproduce a similar statistical significance for detection of treatment

effect in the new region (see, e.g., Hung et al., 1997; Shao and Chow, 2002; Chow et al., 2002; Chow and Shao, 2002a). Therefore, these arguments indicate a fundamental conflict between the evaluation of similarity and the objective of minimizing duplication of clinical data in the ICH E5 guideline.

Consequently, Bayesian methods have been suggested to synthesize the data from both the bridging study and the original region to resolve this conflict (Liu et al., 2002; Shih, 2001). However, some difficulties also arise using the Bayesian method. First, a medicine was approved in the original region due to its substantial evidence of efficacy and safety based on a sufficiently large sample size. The result of the bridging studies using the empirical Bayesian approach will be overwhelmingly dominated by the results of the original region due to an imbalance of sample sizes between the regions. In other words, it is very difficult, if not impossible, to reverse the results observed in the original region; even the result of the bridging study is completely opposite. In addition, the Bayesian method for evaluation of probability for error of decision making on similarity still needs to be worked out. This error probability is extremely crucial for the regulatory authority in the new region to approve a medicine in their jurisdiction.

The ICH E5 guideline provides a rationale for assessment of ethnic factors in the acceptability of foreign data for regulatory strategies of minimizing duplication of clinical data, and it describes the requirements of bridging evidence for extrapolation of foreign clinical data to a new region. It, however, is too premature to develop statistical methods for regulatory implementation unless well-known and scientifically justifiable criteria for (1) the evaluation of sensitivity of medicines to ethnic factors, (2) the assessment of necessity of a bridging study, (3) the determination the nature and type of bridging studies, and (4) the assessment of similarity based on bridging evidence are addressed in the future revision of the ICH E5 guideline.

7.8 VACCINE CLINICAL TRIALS

Similar to clinical development of drug products, there are four phases of clinical trials in vaccine development. Phase I trials are referred to early studies with human subjects. The purpose of phase I trials is to explore the safety and immunogenicity of multiple dose levels of the vaccine under investigation. Phase I trials are usually of a small scale. Phase II trials are to assess the safety, immunogenicity, early efficacy of selected doses of the vaccine, and generate hypotheses for later testing. Phase III trials, which are usually large in scale, are to confirm the efficacy of the vaccine in the target population and/or proving consistency of manufacturing processes. Phase IV trials are usually conducted for collecting additional information regarding long-term safety, immunogenicity, or efficacy of the vaccine to fulfill with regulatory requirement and/or marketing objectives after regulatory approval of the vaccine.

In this section, we provide some design considerations, including special statistical considerations, when conducting vaccine clinical trials. Also included are the classification of types of immunogenicity vaccine trials and statistical methods for various study endpoints.

7.8.1 Basic Design and Statistical Considerations

In its guidance, the FDA discusses the design and statistical considerations for clinical studies to demonstrate the safety, immunogenicity, and efficacy of vaccines (FDA, 1997). These design and statistical considerations are briefly described below.

Safety Studies As indicated in the FDA guidance, studies of comparative safety should be randomized and controlled. Follow-up safety should be actively monitored and prospectively planned with baseline and specific post-vaccination time of assessment. It is suggested that subjects should be actively monitored on designated post-vaccination days for up to one week for most killed and recombinant vaccines or for 14 or more days for most live vaccines. Follow-up should continue through at least 30 days for live or killed vaccines. Thus, as an example, for a particular killed vaccine, subjects should be monitored at 6–12 hours, and days 1, 2, 3, 7, and 30 post-vaccination.

Immunogenicity As the objective of comparative immunogenicity studies is to rule out the important difference between the responses to the study vaccine and a control. Such studies should have sufficient power to rule out clinically meaningful differences in geometric mean titers (GMTs) and/or seroconversion rates. In addition, the study design should take into account the intrinsic variability in assays and subjects. A clinically meaningful difference for each response should be defined prospectively in the clinical protocol.

Efficacy Clinical trials for demonstration of vaccine efficacy should be randomized and controlled. Endpoints used to evaluate efficacy could range from disease incidence to a well-established surrogate marker with activity in correlate of protection. A correlate of protection in vaccine efficacy is generally a laboratory parameter that has been shown from adequate and well-controlled trials to be associated with protection from clinical disease. As indicated in the FDA guidance, an immunological correlate of protection is most useful if clear qualitative and quantitative relationships can be determined, e.g., a certain type and level of antibody correlate with protection.

Statistical Considerations In its guidance, the FDA emphasizes the importance of randomization in vaccine clinical trials. Stratified randomization may be recommended if warranted by the inclusion criteria or known disease risk factors. For nonrandomized trials, the FDA requires the validity and reliability of the evaluation of the study vaccine be provided.

For statistical approaches for data analysis, the FDA indicates that both hypotheses testing and the confidence interval approach are appropriate. The one-sided test for superiority is not an issue for regulatory approval. However, it is suggested that a difference-detection trial be designed for demonstration of superiority of the study vaccine as compared to a control.

For study endpoints, the FDA requires that the evaluation of the study vaccine be performed to rule out a prespecified, clinically meaningful difference between the study vaccine and the control and that should be clearly stated in the study protocol. It is also suggested that the assessment of immune response, common adverse reactions, less common adverse reactions, and rare adverse events should be carefully carried out according to specific statistical approaches as described in the guidance.

For sample size, the FDA requires that sample size calculation for each study endpoint (immunogenicity, safety, and efficacy if applicable) should be performed and the largest sample size should be selected as the one for overall trial enrollment. The FDA indicates that sample size calculation may be performed based on confidence interval rather than on power analysis if it is the planned analysis.

Ellenberg (2001) also discussed some statistical considerations in evaluating the safety of combination vaccines. In their review article, Chan et al. (2003) further posted some

statistical issues when analyzing data from vaccine clinical trials. These statistical issues include intention-to-treat population versus evaluable subset, missing/coarse data and lost-to-follow-up, analysis with stratification variables, and interim analysis. In addition, Chan et al. (2003) suggested that some special considerations, including surrogate markers of activity, the assessment of immune responses, specificity in endpoint definitions of clinical efficacy based on exposure, confidence in safety, and special health economic considerations of the public health impact of vaccination programs be taken into account when planning a vaccine clinical trial.

7.8.2 Types of Vaccine Immunogenicity Trials

Vaccine immunogenicity trials are used to study the immune response to vaccination, which are usually measured by serum antibody concentration or T-cell responses. In early phases of vaccine development, immunogenicity trials are commonly conducted to assess whether the vaccine can induce immunity before entering large efficacy trials. In addition, immunogenicity is often used to determine whether an immune marker to the vaccine can be used as a surrogate or correlate of disease protection. Chan et al. (2003) classified vaccine immunogenicity trials into the following categories of (1) superiority immunogenicity study, (2) dose-response immunogenicity study, (3) consistency lots study, (4) bridging study, (5) combination or multivalent vaccine study, and (6) immunological persistence study, which are briefly outlined below.

Superiority Immunogenicity Study The objective of superiority immunogenicity studies is to assess the superiority of immunogenicity of the vaccine under study as compared to a placebo control. Superiority immunogenicity studies are often conducted in early phases of vaccine development, which can also be performed to claim superiority of one vaccine as compared to another from a different manufacturer with respect to immune responses.

Dose-Response Immunogenicity Study During the development of a new vaccine, dose-response immunogenicity studies are often conducted to assess the immunologic responses across different dose levels of the vaccine. A well-established dose response can help in determining the minimum effective dose and the safe dose. In addition, dose-response studies are useful in studying the kinetic of potency decay and in determining the release and end-expiry dose level and shelf-life of the vaccine.

Consistency Lots Study Before a vaccine can be licensed, the FDA requires that evidence of analytical consistency among at least five lots of vaccine and clinical consistency in at least three lots of vaccine from the same manufacturing process be provided. Frey et al. (1999) suggested that vaccines from three consistency lots and a control vaccine be used for a typical clinical consistency lots study.

Bridging Study When there are minor changes in manufacturing process, storage conditions, routes of administration, or dosing schedules after regulatory approval, the FDA requires a bridging study be conducted to demonstrate that such changes do not have adverse effects on the vaccine effectiveness. An immunogenicity bridging study is usually designed as a noninferiority trial aimed to exclude a clinically significant difference in the immune response between the modified vaccine and the current vaccine.

Combination or Multivalent Vaccine Study A combination vaccine is defined as a vaccine that consists of two or more live organisms, inactivated organisms, or purified antigens combined either by the manufacturer or mixed immediately before administration (FDA, 1997). A combination vaccine is intended to prevent multiple diseases or to prevent one disease caused by different strains or serotypes of the same organism. For establishment of efficacy of a combination vaccine, the immunogenicity of all vaccine components in the combination or multivalent vaccine should be performed to rule out clinically significant differences in immune response rates and/or GMTs between the combined vaccine and the separate but simultaneously administered antigens. In addition, acceptable levels of immunogenicity should be demonstrated for each serotype or component.

Immunological Persistence Study In practice, it is recognized that the duration of vaccine-induced immunity will be considerably longer than the time span of the clinical studies. Thus, immunological persistence study is often conducted to collect data regarding immune responses over multiple years. Life table or time-to-event data analysis is then used to assess the cumulative immunological persistence rate for determination of long-term immunological persistence.

7.8.3 Statistical Methods

As indicated in Chan et al. (2003), one of the most critical steps of evaluations of a new vaccine is to assess the protective efficacy of the vaccine against the target disease. An efficacy trial is often conducted to evaluate whether the vaccine can prevent the disease or reduce the incidence of the disease in the target population. For immunogenicity analysis, the immune response rate and the geometric mean titer or concentration of immune response post-vaccination are usually considered to assess vaccine immunogenicity. In what follows, statistical methods for analysis of these endpoints are briefly outlined.

Disease Incidence/Immune Response

Consider a vaccine clinical trial comparing a new vaccine with a control. Subjects who meet the inclusion/exclusion criteria are randomly assigned to receive either the test vaccine (T) or the placebo control (C). Let p_T and p_C be the true disease incidence rates or immune response rates of the n_T vaccines and n_C controls randomized in the trial, respectively. Thus, the relative reduction in disease incidence for subjects in the vaccine group as compared to the control groups is given by

$$\begin{aligned}\pi &= \frac{p_C - p_T}{p_C} \\ &= 1 - \frac{p_T}{p_C} \\ &= 1 - R.\end{aligned}$$

In most vaccine clinical trials, π has been widely used and is accepted as a primary measure of vaccine efficacy. Note that a vaccine is considered 100% efficacious (i.e., $\pi = 1$) if it prevents the disease completely (i.e., $p_T = 0$). On the other hand, it has no efficacy (i.e., $\pi = 0$) if $p_T = p_C$. Let x_T and x_C be the number of observed diseases for treatment and control groups, respectively. It follows that the natural estimators for p_T and p_C

are given by

$$\hat{p}_T = \frac{x_T}{n_T} \quad \text{and} \quad \hat{p}_C = \frac{x_C}{n_C}.$$

Let $\beta = p_T/p_C$, which can be estimated by

$$\hat{\beta} = \frac{\hat{p}_T}{\hat{p}_C}.$$

By Taylor's expansion and the Central Limit Theorem (CLT), $\hat{\beta}$ is asymptotically distributed as a normal random variable with mean β and variance given by

$$\sigma^2 = \frac{1-p_T}{n_T p_T} + \frac{1-p_C}{n_C p_C}. \quad (7.8.1)$$

For a given confidence level of $1 - \alpha$, the $100 \times (1 - \alpha)$ confidence interval of β is given by

$$(\hat{\beta} - Z(\alpha/2)\hat{\sigma}, \hat{\beta} + Z(\alpha/2)\hat{\sigma}),$$

where $\hat{\sigma}$ is obtained according to (7.8.1) by replacing p_T and p_C by \hat{p}_T and \hat{p}_C , respectively. This leads to the following $100 \times (1 - \alpha)$ confidence interval for π :

$$(1 - \exp(\hat{\beta} + Z(\alpha/2)\hat{\sigma}), 1 - \exp(\hat{\beta} - Z(\alpha/2)\hat{\sigma})).$$

Extremely Low Disease Incidence In many cases, the disease incidence rate is extremely low. In this case, a much larger scale of study is required to demonstrate vaccine efficacy as described in the preceding subsection. For sufficiently large sample sizes and small incidence rates, the numbers of cases in the vaccine groups and the control groups may be approximated by independent Poisson distribution with rate parameters $\lambda_T (\approx n_T p_T)$ and $\lambda_C (\approx n_C p_C)$, respectively. As a result, the number of cases in the vaccine group given the total number of cases (denoted by S) is distributed as a binomial random variable with parameter θ , i.e., $b(S, \theta)$, where

$$\begin{aligned} \theta &= \frac{\lambda_T}{\lambda_C + \lambda_T} = \frac{n_T p_T}{n_T p_T + n_C p_C} \\ &= \frac{R}{R + u} = \frac{1 - \pi}{1 - \pi + u}, \end{aligned}$$

where $u = n_C/n_T$, because θ is a decreasing function in π , testing the hypotheses that

$$H_0: \pi \leq \pi_0 \quad \text{vs.} \quad H_a: \pi > \pi_0$$

is equivalent to testing the following hypotheses:

$$H_0: \theta \geq \theta_0 \quad \text{vs.} \quad H_a: \theta < \theta_0,$$

where

$$\theta_0 = \frac{1 - \pi_0}{1 - \pi_0 + u}.$$

Let x_T and x_C be the number of the observed diseases for the treatment and control, respectively. A natural estimator for θ is given by $\hat{\theta} = x_T/(x_T + x_C)$. The test statistic is given by

$$T = \frac{\sqrt{x_T + x_C}(\hat{\theta} - \theta_0)}{\sqrt{\theta_0(1 - \theta_0)}}$$

Under the null hypothesis, T is asymptotically distributed as a standard normal random variable. Hence, we reject the null hypothesis at α level of significance if $T > Z(\alpha)$.

Relative Vaccine Efficacy In vaccine trials, when the control is a licensed vaccine (an active control), the relative efficacy π can be evaluated through the relative risk (i.e., $R = P_T/P_C$) based on the relationship $\pi = 1 - R$. If the absolute efficacy of the control (i.e., π_C) has been established, one can estimate the absolute efficacy of the test vaccine by

$$\pi_T = 1 - R(1 - \pi_C).$$

For a comparative vaccine trial, it is often designed as a noninferiority trial by testing the following hypotheses:

$$H_0: R \geq R_0 \quad \text{vs.} \quad H_a: R < R_0,$$

where $R_0 > 1$ is a prespecified noninferiority margin or a threshold for relative risk. In practice, the hypotheses regarding relative risk are most often performed based on log-scale. In other words, instead of testing the above hypotheses, we usually consider the following hypotheses:

$$H_0: \log(R) \geq \log(R_0) \quad \text{vs.} \quad H_a: \log(R) < \log(R_0).$$

This becomes the two-sample problem for relative risk.

Composite Efficacy Measure As indicated by Chang et al. (1994), in addition to the prevention of the disease infection, a test vaccine may also reduce the severity of the target disease as well. As a result, it is suggested that a composite efficacy measure be considered to account for both incidence and severity of the disease when evaluating the efficacy of the test vaccine. Chang et al. (1994) proposed the so-called burden-of-illness composite efficacy measure.

Suppose n_T subjects were assigned to receive treatment while n_C subjects were assigned to receive control (placebo). Let x_T and x_C be the number of cases observed in the treatment and the control group, respectively. Without loss of generality, we assume the first x_T subjects in the treatment group and x_C subjects in the control group experienced the events. Let s_{ij} , $i = T, C$; $j = 1, \dots, x_i$ be the severity score associated with the j th case in the i th treatment group. For a fixed $i = T$ or C , it is assumed that s_{ij} are independent and identically distributed random variables with mean μ_i and variance σ_i^2 . Let p_i be the true event rate of

the i th treatment group. The hypotheses of interest is given by

$$H_0: p_T = p_C \text{ and } \mu_T = \mu_C \quad \text{vs.} \quad H_a: p_T \neq p_C \text{ or } \mu_T \neq \mu_C.$$

Let

$$\begin{aligned}\bar{s}_i &= \frac{1}{n_i} \sum_{j=1}^{x_i} s_{ij} \\ \bar{x} &= \frac{n_T \bar{s}_T + n_C \bar{s}_C}{n_T + n_C} \\ s_i^2 &= \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (s_{ij} - \bar{s}_i)^2.\end{aligned}$$

The test statistic is given by

$$T = \frac{\bar{s}_T - \bar{s}_C}{\sqrt{\bar{x}^2 \hat{p}(1 - \hat{p}) (1/n_T + 1/n_C) + \hat{p}(s_0^2/n_T + s_1^2/n_C)}}.$$

Under the null hypothesis, Chang et al., 1994 showed that T is asymptotically distributed as a standard normal random variable. Hence, we would reject the null hypothesis if $|T| > Z(\alpha/2)$.

It should be noted that safety considerations of vaccines are different from that of pharmaceutical products. The methods and measurements chosen for assessment of safety of a vaccine depend on a number of factors, such as the type of vaccine and its specific mechanism for eliciting immune responses. As vaccination can cause allergic or anaphylactic reactions due to the induction of the immune system, it is suggested that all safety variables encountered in a vaccine clinical trial and the analytical approach should be specified in the study protocol. As vaccines will typically be administered to millions of otherwise healthy individuals, it is strongly recommended that rare but serious adverse events be carefully evaluated.

In addition to the statistical methods for analysis of vaccine efficacy and immunogenicity endpoints described above, Durham et al. (1998) considered a nonparametric survival method to estimate the long-term efficacy of a cholera vaccine in the presence of warning protection. For evaluation of long-term vaccine efficacy, as indicated by Chan et al. (2003), the analysis of time-to-event may be useful for determining whether breakthrough rates among vaccinees change over time. It, however, should be noted that sample size calculation may be different depending on the study objectives, the hypotheses of interest, and the corresponding appropriate statistical tests.

Clinical development for vaccine has recently received much attention both from regulatory agencies such as the US. FDA and the pharmaceutical industry. For example, Ellenberg and Dixon (1994) discussed some important statistical issues of vaccine trials (related to HIV vaccine trials). O'Neill (1988b) and Chan and Bohidar (1998) gave asymptotic and exact formulas for sample size and power calculations for vaccine efficacy studies, respectively. Chan et al. (2003) provided a comprehensive review of vaccine clinical trials and statistical issues that are commonly encountered in vaccine clinical trials.

7.9 DISCUSSION

In Section 7.6 we discussed the use of factorial designs for evaluation of combination drugs or combination therapy. Its application, however, should not be limited. Factorial designs are useful in the determination of optimal doses for dose-response studies. For example, for an evaluation of a possible optimal dosage regimen during phase II clinical development, it is necessary to investigate the frequency of drug administrations and dose levels at each administration simultaneously. Therefore the intended trial involves two factors: the frequency of the dosing and the magnitude of dosing. Therefore a factorial design is helpful. Ruberg (1995a, 1995b) considers a full two-factor factorial design for the assessment of the QD and BID regimens and the dose levels given in Table 7.9.1. Ruberg suggests that for the design in Table 7.9.1, a placebo group for the QD regimen might not be needed. In addition, in order to get a stronger comparison between the high dose and placebo of the BID regimen, k in Table 7.9.1 must be greater than 1.

Sometimes the treatment of a certain disease requires inpatient intravenous infusion of one drug as initial therapy followed by outpatient oral administration of another drug as maintenance therapy such as streptokinase and aspirin in the ISIS2 study. Then the factorial design given in Table 7.9.2 can be used in the study of a dose-response relationship. For example, the first row and first column in Table 7.9.2, where placebo of each drug is listed, gives an independent characterization of the pure dose response for different administration, of different drugs.

Although a full factorial design is useful for simultaneously evaluating the joint contribution of different components of a combination drug or therapy, the number of treatment groups can be prohibitively large. This may result in some practical issues during the conduct of the trial. For example, human error is more likely when the study involves a randomization of more than 16 parallel treatment groups for 25 study centers. In addition, due to lack of resource and financial constraints, it may be difficult to conduct a full cross-classification factorial design. A fractional factorial design is an attractive alternative for the evaluation of combination drugs (Box et al. 1978). Although fractional factorial designs are widely found in industry, they are not very common in clinical research. One of the reasons is that the general principle of a fractional factorial design may not be directly applied to meet the objectives for evaluation of combination drugs. Therefore proper modification of the fractional factorial design with valid statistical methods must be made before it can be accepted for assessment of combination therapy. Hung (1996) extended his method to the incomplete factorial design for identification of a combination superior to each component. However, it is not clear how strict superiority can be interpreted if the incomplete factorial design is disconnected as shown in Table 7.9.3 (Searle, 1971).

Table 7.9.1 Factorial Design for Dosing Regimen and Dose Levels

Frequency	Placebo	Dose Levels (Magnitude)		
		1	2	3
QD	n	n	n	n
BID	kn	n	n	kn

Source: Ruberg (1995a).

Table 7.9.2 Factorial Design for Combination of I.V. Infusion Followed by Oral Administration

Dose Level of Oral Administration	I.V. Infusion Rate			
	Placebo	1	2	3
Placebo	<i>n</i>	<i>n</i>	<i>n</i>	<i>n</i>
1	<i>n</i>	<i>n</i>	<i>n</i>	<i>n</i>
2	<i>n</i>	<i>n</i>	<i>n</i>	<i>n</i>
3	<i>n</i>	<i>n</i>	<i>n</i>	<i>n</i>

In combination drug trials, although the primary goal is to identify a combination whose efficacy is better than each component, the quantitative characterization of the contribution made by each component when it is administered as monotherapy is also crucial for clinical practitioners. In addition the interrelationships among the procedures proposed by Hung and others for identification of a superior combination, of dose levels between two drugs, and of each component's effect in monotherapy based on a traditional analysis of variance are quite complex. Hung et al. (1995) propose a two-stage estimation and test procedure for treatment effects of monotherapies in two-by-two factorial trials.

For three-arm trials with combined treatment groups and two component drugs, Laska and Meisner (1989) provide tables of sample sizes for the various powers according to their min procedure which is based on the Student *t*-test and the Wilcoxon rank sum test for univariate clinical endpoints. They found that approximately 25% more patients per treatment group are required for evaluating the superiority of a combination drug than is usual for comparing the difference between two treatment groups. As for the combination trial with a full factorial design, the formula for calculating power is given in Hung et al. (1993), and Hung (1994) can be used to iteratively select sample sizes with a desired power. However, the relative magnitude of the sample sizes required by the procedures proposed by Hung et al. (1993) is not known, whereas the traditional *F*-test is used for finding the treatment difference among dose levels in monotherapy. When the objective is to examine possible interaction between two drugs, as pointed out by Byar and Piantadosi (1985), the sample size for detecting any interaction must be much larger than that for studying the main effects of monotherapy. More research is definitely needed for application to multicenter trials, group sequential methods, and interim analyses of combination drugs.

Since there can be an ethical concern regarding the inclusion of concurrent placebo control in active control equivalence trials, Huque and Dubey (1990) explore the possibility of

Table 7.9.3 An Incomplete Disconnected Factorial Design

Does Level for Drug A	Dose Level of Drug B			
	Placebo	0	1	3
Placebo	×	×		
1	×	×		
2			×	×
3			×	×

Note: The dose × indicates the combination being studied.

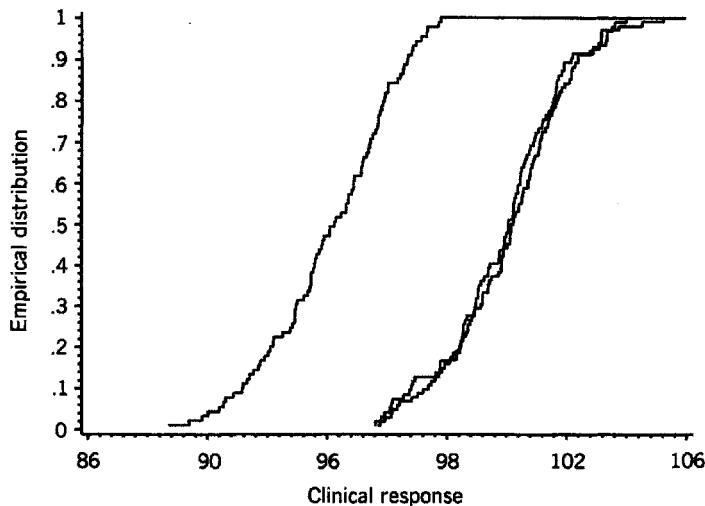


Figure 7.9.1 Empirical cumulative distribution of an active control equivalence trial with a concurrent placebo control.

designing a group sequential trial by performing interim analyses for the purpose of stopping the placebo arm but continuing the test and active reference control. However, in order to stop the placebo group early, based on results from interim analyses, the randomization codes of all three groups must be broken to reveal the treatment assignment of each patient. As a result bias will be unavoidably introduced into the subsequent assessment between the test and active control, and consequently the primary goal for an unbiased evaluation of the equivalence between the two drugs may be in jeopardy.

For binary data the sample size determination and equivalence limits depend on not only the assumed reference response rate but also the overall response rate (Huque and Dubey, 1990). It should be noted that the observed reference and overall response rates might differ from the assumed rates used for sample size determination at the planning stage of the trial. In addition they may be different from one interim examination to another. It is, however, not clear whether the same prespecified equivalence limit should be used for all interim analyses. Durrleman and Simon (1990) and Jennison and Turnbull (1993) offer some useful thoughts on the design, conduct, monitoring and analysis of sequential active control equivalence trials. However, their methodology does not include a joint evaluation of equivalence and superiority in the efficacy of active drugs in relation to the placebo (Morikawa and Yoshida, 1995).

Note that our discussion in Section 7.4 only focused on the average equivalence rather than the population equivalence. Figure 7.9.1 illustrates the case where the cumulative distribution of the test and reference active control are equivalent, assuming that they both are (stochastically) larger than that of the placebo group. For assessment of equivalence with inclusion of a placebo concurrent control in cumulative distribution, the hypotheses are given as.

$$\begin{aligned}
 H_0: & |F_T - F_R| \geq d \quad \text{or} \quad F_T \leq F_P \quad \text{or} \quad F_R \leq F_P, \\
 \text{vs. } H_a: & |F_T - F_R| < d \quad \text{and} \quad F_T > F_P \quad \text{and} \quad F_R > F_P,
 \end{aligned} \tag{7.9.1}$$

where F_T , F_R , and F_P are the cumulative distribution functions (CDF) of the clinical responses of test, reference active control, and placebo concurrent control, respectively. Hypotheses (7.9.1) can be further decomposed into three sets.

$$\begin{aligned} H_{0E}: |F_T - F_R| &\geq d, \\ \text{vs. } H_{aE}: |F_T - F_R| &< d, \end{aligned} \quad (7.9.2)$$

$$\begin{aligned} H_{0T}: F_T &\leq F_P, \\ \text{vs. } H_{aT}: F_T &> F_P, \end{aligned} \quad (7.9.3)$$

and

$$\begin{aligned} H_{0R}: F_R &\leq F_P, \\ \text{vs. } H_{aR}: F_R &> F_P. \end{aligned} \quad (7.9.4)$$

The standard one-sided Kolmogorov-Smirnov test (Conover, 1980) can be used to verify whether the distributions of both the test and reference active control are (stochastically) larger than those of the placebo. Recently, as an alternative, Wellek (1993) proposed a method that tests the equivalence between two cumulative distributions. However, his procedure requires an estimation of a noncentrality parameter of the chi-square distribution under hypotheses (7.9.2). More research on equivalence in distributions is needed.

8

ANALYSIS OF CONTINUOUS DATA

8.1 INTRODUCTION

As was pointed out in Chapter 1, a well-designed protocol can ensure the success of clinical research. A well-designed protocol focuses on all details including how the intended clinical trial is to be carried out and how the data are to be collected. It also must reflect Section 312.23 of 21 CFR which requires that a statistical section, which describes how the data are to be analyzed, must be a part of the protocol. It is equally important to implement good clinical practice/good statistical practice to provide an accurate and reliable assessment of the efficacy and safety of the test drug under study. For this purpose appropriate and valid statistical methods must be determined for the collected data. The collected clinical data are referred to as the responses of the clinical endpoints or variables of patients under study. These clinical endpoints or variables are used to measure or evaluate the characteristics of the test drug product for treatment of patients with certain diseases under study. Basically the collected clinical data can be classified as either qualitative (categorical data) or quantitatively (numerical or measurement data).

When the characteristic concerns a qualitative trait that is only classified in categories and not numerically measured, the resulting data are called categorical data. For the categorical data, subjects are placed in the proper category or group, and the number of subjects in each category is enumerated. Each subject must be fit into exactly one category. In clinical trials gender and race are typical categorical variables. Note that if there is order among the categories, the resulting data are called ranked data or ordered categorical data. For example, the severity and intensity of pain are considered ranked data. For ranked data there are not necessarily equal intervals or differences between ranks. On the other hand, if the characteristic is measured on numerical scales, then the resulting data are numerical

data. Note that if the numerical scales are made of distinct numbers with gaps in between such as the number of drinks daily, then the variables are referred to as discrete variables. Clinical endpoints are said to be continuous if they can take any values in an interval. Some variables such as height, weight, survival time, and diastolic and systolic blood pressures are continuous variables. The responses of a continuous variable are considered continuous data.

In general, different statistical methods are usually applied for different types of data. In this chapter, our primary emphasis will be placed on statistical analysis for continuous data. Analysis for categorical data will be discussed in the next chapter. In the next section the concept of estimation including point and interval estimates will be briefly introduced. Statistical tests such as paired t and two sample t are given in Section 8.3. In Sections 8.4 and 8.5 the method of variance and covariance analysis are discussed. The use of nonparametric methods and the application of repeated measures are described in Sections 8.6 and 8.7 respectively. A brief discussion is presented in the last section.

8.2 ESTIMATION

As mentioned, clinical endpoints are used to measure or describe the characteristics of a test drug for the treatment of patients with certain diseases. Based on the observed responses of the clinical endpoints, we can draw some statistical inferences on the characteristics of the test drug. For example, reductions in seated diastolic and systolic blood pressures are often used to evaluate the effect of a test drug in the treatment of patients with mild to moderate hypertension. In this section we will focus on the estimation of some quantities such as the mean and variability of the characteristics.

Suppose that there are N subjects in a targeted patient population. For a given clinical endpoint, let y_1, y_2, \dots, y_N be the true responses of the clinical endpoint for the N subjects in the targeted patient population. Then, the population mean and population variance of the clinical endpoint are given by

$$\mu = \frac{1}{N} \sum_{i=1}^N y_i,$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2,$$

respectively. Note that μ and σ^2 are usually unknown. In practice, we make a statistical inference on μ to assess the efficacy of the drug under study. For example, μ may represent the mean reduction in the seated diastolic blood pressure in patients with mild to moderate hypertension. As discussed in Chapter 4, in practice, since N can be very large (e.g., a few million people), it is impossible to observe all the $y_i, i = 1, \dots, N$ values. Therefore, as an alternative, we select a random sample of size n and then draw statistical inference on μ from an analysis of information contained in the sample data provided that the random sample is a representative of the targeted patient population. Intuitively we can use the mean and standard deviation of the n observations from the random sample to estimate μ and σ . That is, we let $\{Y_1, Y_2, \dots, Y_n\}$ be a random sample of size n selected from the targeted patient population $\{y_1, y_2, \dots, y_N\}$, then the realization of $\{Y_1, Y_2, \dots, Y_n\}$ is a subset of $\{y_1, y_2, \dots, y_N\}$, where $2 \leq n \leq N$. Then μ and σ can be estimated by the sample

mean and sample standard deviation, which are given by

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i,$$

$$s = \left\{ \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 \right\}^{1/2},$$

respectively. When the sample size n is sufficiently large, \bar{Y} and s would be quite close to the unknown population mean μ and population standard deviation σ , respectively.

Note that if $\{Y_1, Y_2, \dots, Y_n\}$ results from independent selections, then each has the same distribution as the population. Since random samples vary, the values of \bar{Y} also vary. In other words, \bar{Y} is also a random variable. The sampling distribution of \bar{Y} can be obtained by the following steps:

1. List all possible samples of size n .
2. Calculate the values of \bar{Y} for each sample.
3. List the distinct values of \bar{Y} , and calculate the corresponding probabilities by identifying all of the samples that yield the same value of \bar{Y} .

For the selection of a random sample of size n , there are a total of M possible selections for a population of size N , where

$$M = \binom{N}{n} = \frac{N(N-1) \cdots (N-n)}{n!}.$$

Let \bar{Y}_k be the mean of the k th random sample, where $k = 1, \dots, M$. Then, the expectation of \bar{Y} is the average of the sample means over all possible samples,

$$E(\bar{Y}) = \frac{1}{M} \sum_{k=1}^M \bar{Y}_k.$$

If

$$E(\bar{Y}) = \mu,$$

then \bar{Y} is said to be an unbiased estimate of μ . Similarly, the standard error of \bar{Y} can be obtained as follows:

$$\sigma_{\bar{Y}} = SD(\bar{Y}) = \sqrt{\text{var}(\bar{Y})} = \left\{ \frac{1}{M} \sum_{k=1}^M [\bar{Y}_k - E(\bar{Y})]^2 \right\}^{1/2}.$$

Since \bar{Y} is the average of n random variables with common mean μ and variance σ^2 , it can be verified that

$$\sigma_{\bar{Y}}^2 = \frac{1}{n} \sigma^2.$$

Note that if the population is normal with mean μ and standard deviation σ , then \bar{Y} has the normal distribution with mean μ and standard deviation σ/\sqrt{n} . However, if the population is an arbitrary population with mean μ and standard deviation σ , then \bar{Y} is approximately normal with mean μ and standard deviation σ/\sqrt{n} when n is large (e.g., $n \geq 30$). As a result, since σ can be consistently estimated by s ,

$$z = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

is approximately distributed as a standard normal with mean 0 and standard deviation 1. This property is known as the *Central Limit Theorem*. As a result, when n is large, in addition to the point estimate, we can provide an interval estimate for μ based on the Central Limit Theorem. A large samples $100(1 - \alpha)\%$ confidence interval for μ is given by

$$\left(\bar{Y} - \frac{Z(\alpha/2)s}{\sqrt{n}}, \bar{Y} + \frac{Z(\alpha/2)s}{\sqrt{n}}\right), \quad (8.2.1)$$

where $Z(\alpha/2)$ denotes the upper $(\alpha/2)$ th quantile of the standard normal distribution (i.e., the area to the right of $Z(\alpha/2)$ is $\alpha/2$). The $100(1 - \alpha)\%$ confidence interval is a random interval that will cover the true μ with probability $(1 - \alpha)$. In other words, if random samples are repeatedly drawn from the same population and a $100(1 - \alpha)\%$ confidence interval is calculated from each sample, then about $100(1 - \alpha)\%$ of these intervals will contain the true but unknown value μ . In practice, however, since we only draw one sample (i.e., conduct the intended clinical trial once), we never know what happens in a single application. Therefore, when we say that *we are $100(1 - \alpha)\%$ confident that the confidence interval obtained from the sample will cover the true μ* , our confidence draws from the success rate of $100(1 - \alpha)\%$ in many applications.

Note that the above interval estimate is derived based on the Central Limit Theorem with large n . In practice, n may not be large, especially when the trial involves enormous medical expenditures or the disease is rare. In this case, the question *What is the sample distribution of \bar{Y} when n is not large?* needs to be addressed in order to provide an interval estimate for μ . As indicated earlier, when \bar{Y} is based on a random sample of size n from a normal population with mean μ and standard deviation σ , \bar{Y} is distributed as a normal distribution with mean μ and standard deviation σ/\sqrt{n} . Since σ is usually unknown, an intuitive approach is to estimate σ by the sample standard deviation s . This leads to the following t statistic

$$t = \frac{\bar{Y} - \mu}{s/\sqrt{n}}.$$

The distribution of the above t statistic is known as Student's t distribution with $n - 1$ degrees of freedom. As discussed above, the Student t distribution gets closer to the standard normal as n becomes larger. For small n , although the t distributions are also symmetric about 0, they have tails that are more spread out than the standard normal distribution. Therefore, based on the t statistic, we can construct a $100(1 - \alpha)\%$ confidence interval for μ for small n as follows:

$$\left(\bar{Y} - \frac{t(\alpha/2, n - 1)s}{\sqrt{n}}, \bar{Y} + \frac{t(\alpha/2, n - 1)s}{\sqrt{n}}\right), \quad (8.2.2)$$

where $t(\alpha/2, n - 1)$ is the upper $(\alpha/2)$ th quantile of the t distribution with $n - 1$ degrees of freedom.

To illustrate the estimation procedures described above, consider a clinical trial conducted in order to evaluate the safety of an injectable dosage form of a drug product (denoted by treatment A) that is compared to a placebo (denoted by treatment B) in subjects undergoing stress echocardiography. This trial was a multicenter, single-blind, and randomized study. The primary safety variables included incidence of adverse events and changes in laboratory parameters. Table 8.2.1 lists the partial data of pre- and post-treatment platelet counts of 30 patients from the five centers. Table 8.2.2 provides the summary statistics of the platelet counts. From Table 8.2.2 and using (8.2.2), we obtain the confidence intervals

Table 8.2.1 Pre- and Post-Treatment Platelet Counts

Center	Subject	Pre-treatment	Post-treatment	Treatment
1	1	359,000	396,000	B
1	2	200,000	184,000	A
1	3	149,000	151,000	A
1	4	235,000	242,000	B
1	5	174,000	177,000	B
1	6	271,000	203,000	A
2	7	180,000	199,000	B
2	8	252,000	256,000	A
2	9	188,000	187,000	B
2	10	211,000	210,000	A
2	11	217,000	199,000	B
2	12	195,000	215,000	A
3	13	266,000	192,000	B
3	14	267,000	205,000	A
3	15	217,000	233,000	A
3	16	247,000	241,000	B
3	17	204,000	188,000	A
3	18	340,000	302,000	B
4	19	175,000	173,000	A
4	20	242,000	233,000	B
4	21	383,000	389,000	A
4	22	165,000	154,000	A
4	23	213,000	226,000	B
4	24	244,000	241,000	B
5	25	243,000	223,000	A
5	26	205,000	167,000	B
5	27	373,000	315,000	B
5	28	252,000	271,000	A
5	29	291,000	250,000	B
5	30	162,000	146,000	A

Table 8.2.2 Summary Statistics of Pre- and Post-Treatment Platelet Counts

Treatment	N	Pre-treatment	Post-treatment	Difference
A	15	223.1 (15.2)	213.4 (15.7)	-9.7 (6.7)
B	15	251.6 (16.4)	237.8 (15.7)	-13.8 (7.9)
A-B	15	-28.5 (15.8)	-24.4 (15.7)	4.1 (7.3)

Note: The values are corresponding platelet counts ($\times 10^3/\mu\text{L}$) and the numbers in the parentheses, are standard errors.

Table 8.2.3 Confidence Intervals Based on *t*-Statistic

Treatment	95% Confidence Interval		
	Pre-treatment	Post-treatment	Difference
<i>A</i>	(214.7, 231.5)	(204.7, 222.1)	(-13.4, -6.0)
<i>B</i>	(242.5, 260.7)	(229.1, 246.5)	(-18.1, -9.5)
<i>A</i> - <i>B</i>	(-37.5, -19.5)	(-33.4, -15.4)	(-9.6, 1.3)

Note: The values in the parentheses are corresponding platelet counts ($\times 10^3/\mu\text{L}$).

for the average platelet counts at the baseline (pre-treatment) and at endpoints (post-treatment) for both treatments. For example, for treatment *A*, since

$$\bar{Y}_A = 223066.67 \quad \text{and} \quad s_A = 15219.18,$$

the 95% confidence interval for the average pre-treatment platelet counts is given by

$$\begin{aligned}\bar{Y}_A \pm \frac{t(\alpha/2, n-1)s_A}{\sqrt{n}} &= 223066.67 \pm (2.14)\left(\frac{15219.18}{\sqrt{15}}\right) \\ &= 223066.67 \pm 8409.29 \\ &= (214657.38, 231475.96).\end{aligned}$$

Similarly, a 95% confidence interval for the average post-treatment platelet counts can be obtained. Table 8.2.3 gives 95% confidence intervals for the average pre- and post-treatment platelet counts for both treatments.

8.3 TEST STATISTICS

In clinical trials, as pointed out earlier, a typical approach for assessment of the efficacy of a test drug is first to demonstrate that the mean effect (i.e., μ) is significantly different from zero or a prespecified small value (e.g., μ_0) and then claim that the test drug is effective by showing that there is a desired power for detection of a clinically meaningful difference. Therefore, the hypotheses of interest are given by

$$\begin{aligned}H_0: \mu &= \mu_0, \\ \text{vs. } H_a: \mu &\neq \mu_0.\end{aligned}\tag{8.3.1}$$

If the above hypotheses concern the mean of a normal population, then we may consider the following test statistic for testing (8.3.1):

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}},$$

which is distributed as the Student *t* distribution with $n - 1$ degrees of freedom under H_0 . Note that the alternative hypothesis given in (8.3.1) is a two-sided hypothesis. In

practice, as mentioned in Chapter 2, we can consider either one of the following one-sided hypotheses:

$$H_a: \mu > \mu_0, \quad (8.3.2)$$

or

$$H_a: \mu < \mu_0.$$

We would reject the null hypothesis of (8.3.1) at the α level of significance if

$$|t| \geq t(\alpha/2, n - 1)$$

On the other hand, for the one-sided hypotheses, we can reject the null hypothesis at the α level of significance if

$$t \geq t(\alpha, n - 1) \quad \text{or} \quad t \leq -t(\alpha, n - 1)$$

for alternative hypotheses $H_a: \mu > \mu_0$ or $H_a: \mu < \mu_0$, respectively.

Note that the concepts of hypotheses testing and confidence interval are operationally equivalent. For testing the hypotheses in (8.3.1), the rejection region of the level α test is

$$\left| \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} \right| \geq t(\alpha/2, n - 1).$$

Consequently, the acceptance region is given by

$$-t(\alpha/2, n - 1) < \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} < t(\alpha/2, n - 1),$$

which can be rewritten as

$$\bar{Y} - \frac{t(\alpha/2, n - 1)s}{\sqrt{n}} < \mu_0 < \bar{Y} + \frac{t(\alpha/2, n - 1)s}{\sqrt{n}}. \quad (8.3.3)$$

Note that the above expression of the acceptance region is the same as the confidence interval for μ given in (8.2.2). Expression (8.3.3) indicates that any given null hypothesis μ_0 will not be rejected at the significance level of α if μ_0 lies within the $100(1 - \alpha)\%$ confidence interval of μ . Therefore, having established a $100(1 - \alpha)\%$ confidence interval for μ , all null hypotheses for values of μ_0 lying outside this interval will be rejected at the α level of significance and all those lying inside will not be rejected. As indicated in Chapter 2, an observed p -value is always reported in accordance with the test. We would reject the null hypothesis if the p -value is less than the α level of significance. The confidence interval approach is regarded as a more comprehensive inference procedure than testing a single null hypothesis because a confidence interval statement in effect tests many null hypotheses at the same time.

Paired t Test

For some clinical trials such as noncomparative open-label studies, one of the primary objectives is to evaluate the effect before and after the treatment based on changes from baseline of the endpoint. In this situation, a paired- t test is useful for within treatment comparison. Let

Y_{Bi} and Y_{Ei} be the responses for the i th subject at baseline (or pre-treatment) and endpoint (or post-treatment), respectively. Statistically, each subject is considered as a block. Therefore, although (Y_{Bi}, Y_{Ei}) are independent of one another, Y_{Bi} and Y_{Ei} within the i th subject are dependent. Consider the endpoint changes from baseline as follows:

$$D_i = Y_{Ei} - Y_{Bi}$$

where $i = 1, \dots, n$. Since $D_i = Y_{Ei} - Y_{Bi}$, $i = 1, \dots, n$, remove the block effect, it is reasonable to assume that they constitute a random sample from a population with mean μ_d and standard deviation σ_d , where μ_d represents the mean difference of before and after treatment. In other words,

$$\begin{aligned}\mu_d &= E(D_i), \\ \sigma_d &= \sqrt{\text{var}(D_i)}, \quad i = 1, \dots, n.\end{aligned}\tag{8.3.4}$$

The hypotheses of interest for a within treatment comparison are

$$\begin{aligned}H_0: \mu_d &= \mu_0, \\ \text{vs. } H_a: \mu_d &\neq \mu_0.\end{aligned}\tag{8.3.5}$$

We reject the null hypothesis at the α level of significance if

$$|t| = \left| \frac{\bar{D} - \mu_0}{s_d / \sqrt{n}} \right| \geq t(\alpha/2, n - 1),\tag{8.3.6}$$

where

$$\begin{aligned}\bar{D} &= \frac{1}{n} \sum_{i=1}^n D_i, \\ s_d &= \sqrt{\frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2},\end{aligned}$$

and $t(\alpha/2, n - 1)$ is the upper $(\alpha/2)$ th quantile of a t distribution with $n - 1$ degrees of freedom. Similarly, we can reject the null hypothesis at the α level of significance if

$$t \geq t(\alpha, n - 1) \quad \text{or} \quad t \leq -t(\alpha, n - 1)$$

for alternative hypotheses $H_a: \mu_d > \mu_0$ or $H_a: \mu_d < \mu_0$, respectively.

Note that the above test for hypotheses given (8.3.5) is called a paired- t test; the paired- t test is commonly used for a within-treatment comparison between the baseline and endpoint of a clinical trial. The corresponding confidence interval for μ_d is given by

$$\left(\bar{D} - \frac{t(\alpha/2, n - 1)s_d}{\sqrt{n}}, \bar{D} + \frac{t(\alpha/2, n - 1)s_d}{\sqrt{n}} \right).$$

To illustrate the application of a paired- t test, consider the example of platelet counts given in the previous section. Based on difference in platelet counts between pretreatment

and post-treatment, paired-*t* tests for treatment *A* and treatment *B* are given by

$$|t_A| = \left| \frac{\bar{D}}{s_d/\sqrt{n}} \right| = \left| \frac{-9666.67}{6702.29/\sqrt{15}} \right| = 5.59$$

and

$$|t_B| = \left| \frac{\bar{D}}{s_d/\sqrt{n}} \right| = \left| \frac{-13800.00}{7855.97/\sqrt{15}} \right| = 6.80,$$

respectively. Since t_A and t_B are both greater than $t(0.025, 14) = 2.14$, we reject the null hypothesis of no difference between pretreatment and post-treatment average platelet counts for both treatments at the α level of significance. Note that 95% confidence intervals of μ_d for treatments *A* and *B*, as given in Table 8.2.3, are $(-13.4, -6.0)$ and $(-18.1, -9.5)$ ($10^3/\mu\text{L}$), respectively, which do not contain 0. We can conclude that there is a significant difference between the pretreatment and post-treatment platelet counts for both treatments. This result is consistent with that obtained from the paired-*t* tests.

Recall that Table 8.2.1 showed that the pre- and post-treatment platelet counts are positively correlated. Thus, an analysis of changes from the baseline (i.e., post-treatment minus pretreatment) is more appropriate because it reduces the variability. As indicated in Table 8.2.2, the variability is reduced by almost a half. In addition a direct comparison suggests that the average platelet counts of treatment *B* are significantly greater than those of treatment *A*. However, Table 8.2.3 indicates that the mean decrease in platelet counts for treatment *A* is not statistically different from that of treatment *B*.

Two-Sample *t* Test

In comparative clinical trials, the primary objective is to evaluate the efficacy and safety of a test drug as compared to a control (e.g., a placebo control or an active control). For this purpose a parallel-group design in which patients are randomly assigned to receive either the test drug or the control is usually employed. In this design setting we basically compare two treatment groups (the test drug and the control) or two populations on the basis of independent samples (i.e., patients who are enrolled in the two treatment groups). Let Y_{ij} , $i = 1, 2$ and $j = 1, \dots, n_i$, be random samples of sizes n_1 and n_2 from population 1 (i.e., patients under treatment of the test drug) and population 2 (i.e., patients under treatment of the control), respectively. Note that the two samples are independent. In other words, the responses under one treatment are uncorrelated to the responses under the other treatment. Suppose that the two populations are normal with means μ_1 and μ_2 and standard deviations σ_1 and σ_2 , respectively. Then, the efficacy of the test drug, as compared to the control, can be examined by testing the following hypotheses:

$$\begin{aligned} H_0: \mu_1 &= \mu_2, \\ \text{vs. } H_a: \mu_1 &\neq \mu_2. \end{aligned} \tag{8.3.7}$$

When sample sizes n_1 and n_2 are large, according to the central limit theorem, the statistic

$$Z = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

is approximately normal with mean 0 and standard deviation 1, where \bar{Y}_1 and \bar{Y}_2 and s_1^2 and s_2^2 are the sample means and sample variances of the two samples respectively. Therefore, we can reject the null hypothesis at the α level of significance if

$$|Z| \geq Z(\alpha/2),$$

where $Z(\alpha/2)$ is the upper $(\alpha/2)$ th quantile of the standard normal.

When sample sizes n_1 and n_2 are small, the above test is not valid. However, under the assumption that (1) both populations are normal and (2) $\sigma_1 = \sigma_2 = \sigma$, the test statistic

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (8.3.8)$$

has Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom, where

$$s = \left\{ \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right\}^{1/2},$$

which is a pooled estimate of the common standard deviation σ . Therefore, we can reject the null hypothesis of (8.3.7) at the α level of significance if

$$|t| \geq t(\alpha/2, n_1 + n_2 - 2),$$

where $t(\alpha/2, n_1 + n_2 - 2)$ is the upper $(\alpha/2)$ th quantile of a t distribution with $n_1 + n_2 - 2$ degrees of freedom. Similarly, we can reject the null hypothesis at the α level of significance if

$$t \geq t(\alpha, n_1 + n_2 - 2) \quad \text{or} \quad t \leq -t(\alpha, n_1 + n_2 - 2)$$

for alternative hypotheses $H_a: \mu_1 > \mu_2$ or $H_a: \mu_1 < \mu_2$, respectively. Note that the above test is known as a two-sample t test. Based on this test, a $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{Y}_1 - \bar{Y}_2) \pm t(\alpha/2, n_1 + n_2 - 2) s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

In practice, σ_1 is never the same as σ_2 . Whether or not to pool the data for an estimate of σ (under the assumption that $\sigma_1 = \sigma_2 = \sigma$) affects the validity of the test statistic described in (8.3.8). The relative magnitude of the two sample variances s_1^2 and s_2^2 is usually the indicator used to determine whether or not to pool. For example, if s_1^2/s_2^2 is far away from 1, then the assumption that $\sigma_1 = \sigma_2 = \sigma$ will be in doubt. In practice, as a rule of thumb, if the ratio falls within the range of

$$\frac{1}{4} \leq \frac{s_1^2}{s_2^2} \leq 4,$$

then we can pool the data for an estimate of the common standard deviation σ .

In the case where $\sigma_1 \neq \sigma_2$, we may reject the null hypothesis if

$$|t^*| \geq t(\alpha/2, v^*),$$

where

$$t^* = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{w_1 + w_2}}, \quad (8.3.9)$$

and

$$\begin{aligned} v^* &= \frac{(w_1 + w_2)^2}{(w_1^2/(n_1 - 1) + w_2^2/(n_2 - 1))}, \\ w_1 &= s_1^2/n_1, \quad \text{and} \quad w_2 = s_2^2/n_2. \end{aligned}$$

Based on t^* , a $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ can be obtained as follows:

$$\bar{Y}_1 - \bar{Y}_2 \pm t(\alpha/2, v^*)\sqrt{w_1 + w_2}, \quad (8.3.10)$$

The above confidence interval is known as Satterthwaite's confidence interval.

Again, we consider the example of platelet counts to illustrate the use of a two-sample t test. The differences in platelet counts between pretreatment and post-treatment for the two treatments concern two independent samples. Therefore, we can use test statistic described in (8.3.8) or (8.3.9) to test the null hypothesis of no treatment difference. As can be seen from Table 8.2.2,

$$s_A = 6702.29 \quad \text{and} \quad s_B = 7855.97.$$

Since

$$\frac{s_A^2}{s_B^2} = 0.729,$$

which is within the range $(0.25, 4)$, we can apply the test statistic in (8.3.8) to test the null hypothesis of no treatment difference as follows:

$$\begin{aligned} |t| &= \left| \frac{(\bar{Y}_A - \bar{Y}_B)}{s \sqrt{1/n_A + 1/n_B}} \right| \\ &= \left| \frac{(-9666.67) - (-13800.00)}{\sqrt{7109121.69}} \right| \\ &= \left| \frac{4133.33}{2666.29} \right| = 1.55, \end{aligned}$$

which is less than $t(0.025, 8) = 2.048$. Therefore, we fail to reject the null hypothesis of no treatment difference at the $\alpha = 5\%$ level of significance. From (8.3.8) the corresponding 95% confidence interval can also be obtained, which is given by $(-9593.89, 1327.23)$. Since the 95% confidence interval contains 0, we reach the same conclusion that there is no significant difference between the two treatments. Note that since $n_A = n_B = 15$, the test statistics in (8.3.8) and (8.3.9) are essentially the same. Both tests fail to reject the null hypothesis of no treatment difference.

8.4 ANALYSIS OF VARIANCE

In previous sections we discussed the estimations and tests for comparing two treatments. Suppose that we are interested in evaluating the efficacy of several test drugs for the same indication compared to a placebo control. Although we may conduct a number of separate clinical trials whereby each compares a test drug with the placebo control, it is more efficient to conduct one clinical trial and to simultaneously compare the test drugs and the placebo. In this case the two-sample t test and its corresponding confidence interval for mean differences is not appropriate. The alternative analysis of variance (ANOVA) method can be applied to compare several population means.

One-Way Classification

Let Y_{ij} be the response of the j th subject on treatment i , where $j = 1, \dots, n_i$ and $i = 1, \dots, k$. Also let μ_i denote the mean of the i th treatment. Then, the hypotheses for a simultaneous evaluation of the k treatments is given by

$$\begin{aligned} H_0: \mu_1 &= \mu_2 = \dots = \mu_k, \\ \text{vs. } H_a: \mu_i &\neq \mu_j \quad \text{for some } 1 \leq i \neq j \leq k. \end{aligned} \quad (8.4.1)$$

To derive a test for the above hypotheses, consider the following model:

$$\begin{aligned} Y_{ij} &= \mu_i + e_{ij} \\ &= \mu + \tau_i + e_{ij}, \quad j = 1, \dots, n_i, i = 1, \dots, k, \end{aligned} \quad (8.4.2)$$

where μ is the overall mean, τ_i denotes the effect of the i th treatment, and e_{ij} are independent random errors in observing Y_{ij} with mean 0 and standard deviation σ . The analysis of variance concept is to partition the observations Y_{ij} into contributions from different sources. Let $\bar{Y}_{..}$ and $\bar{Y}_{i..}$ be the overall sample mean and the sample mean for the i th treatment group, respectively. Then, the deviation of an individual observation from the overall sample mean can be described as

$$Y_{ij} - \bar{Y}_{..} = (\bar{Y}_{i..} - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_{i..}) \quad (8.4.3)$$

or

$$Y_{ij} = \bar{Y}_{..} + (\bar{Y}_{i..} - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_{i..}).$$

As a result the deviation of an individual observation from the overall sample mean consists of two components: the sum of differences among the means of the treatments and the random variation in measurements within the same treatment. Expression (8.4.3) leads to the following partitions of sum of squares (SS):

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^k n_i (\bar{Y}_{i..} - \bar{Y}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i..})^2.$$

The first term on the left-hand side is the total sum of squares, and the first and second terms on the right-hand side are sum of squares due to treatment (denoted by SSA) and

sum of squares due to error (denoted by SSE). Therefore we have

$$\text{SST} = \text{SSA} + \text{SSE}.$$

The mean square due to treatment and error is defined as

$$\begin{aligned}\text{MSA} &= \frac{\text{SSA}}{k - 1} \\ &= \frac{1}{k - 1} \sum_{i=1}^k n_i (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2, \\ \text{MSE} &= \frac{\text{SSE}}{\sum_{i=1}^k (n_i - 1)} = \frac{1}{\sum_{i=1}^k (n_i - 1)} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2.\end{aligned}$$

Under the null hypothesis, the population means are all equal (i.e., $\mu_i = \mu$ for all i or $\tau_i = 0$ for all i), $\bar{Y}_{i\cdot} - \bar{Y}_{..}$ is expected to be small, and consequently its mean square (MSA) is expected to be small. On the other hand, if the population means differ, then MSA is likely to be large. However, the question that is not clear is *How large a treatment mean will yield a statistically significant difference?* To determine whether the population means are statistically significantly different from one another, since the mean square error provides an estimate of σ^2 , we consider the ratio between the treatment mean square and the mean sum of squares due to error. Thus, under the null hypothesis of (8.4.1), the ratio

$$\begin{aligned}F &= \frac{\text{MSA}}{\text{MSE}} \\ &= \frac{\text{SSA}/(k - 1)}{\text{SSE}/(N - k)}\end{aligned}$$

has an F distribution with $(k - 1, N - k)$ degrees of freedom, where $N = \sum_{i=1}^k n_i$. Therefore we can reject the null hypothesis at the α level of significance if

$$F \geq F(\alpha, k - 1, N - k),$$

where $F(\alpha, v_1, v_2)$ is the upper α th quantile of the F distribution with (v_1, v_2) degrees of freedom. Table 8.4.1 provides an analysis of variance table for comparing k treatments.

Simultaneous Confidence Intervals

Suppose that we reject the null hypothesis that there are no differences among the k treatments. Then, we need to show that there is significant difference among the k treatments. In this case it is often of interest to construct a confidence interval for $\mu_i - \mu_{i'}$ for $i \neq i'$ to determine which treatments have unequal means. Based on the mean square error from the analysis of variance, a $100(1 - \alpha)\%$ confidence interval for $\mu_i - \mu_{i'}$ where $i \neq i'$ can be constructed as follows:

$$\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot} \pm t(\alpha/2, N - k)s \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}, \quad (8.4.4)$$

Table 8.4.1 ANOVA Table for Comparing k Treatments

Source of Variation	Sum of Squares	df	Mean Squares
Treatment	$SSA = \sum_{i=1}^k n_i(\bar{Y}_i - \bar{Y}_{..})^2$	$k-1$	$MSA = \frac{SSA}{k-1}$
Error	$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$	$N-k$	$MSE = \frac{SSE}{N-k}$
Total	$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2$	$N-1$	

Note: $N = \sum_{i=1}^k n_i$.

where

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{N-k}}$$

and $t(\alpha/2, N-k)$ is the upper $(\alpha/2)$ th quantile of t distribution with $N-k$ degrees of freedom. In practice, one can consider comparing means pairwisely according to (8.4.4) if the overall F test shows significant differences in the means. However, for k treatments there is a total of

$$m = \binom{k}{2} = \frac{k(k-1)}{2}$$

pairwise differences $\mu_i - \mu_{i'}$, and (8.4.4) applied to all pairs yield m confidence statements in which each as a $100(1-\alpha)\%$ level of confidence. Therefore, it is difficult to determine what level of confidence will be achieved for claiming that these m statements are all correct. To overcome this problem, several procedures have been developed in such a manner that the joint probability that all the statements are true is guaranteed not exceed a predetermined level such as the α level of significance. The most commonly used method among these procedures probably is Bonferroni's simultaneous intervals, also known as multiple- t confidence intervals since the α level is adjusted for multiple comparisons among means. The $100(1-\alpha)\%$ Bonferroni's simultaneous intervals for the m pairwise difference $\mu_i - \mu_{i'}$ are given by

$$\bar{Y}_i - \bar{Y}_{i'} \pm t(\alpha/2m, N-k)s \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}},$$

where $s = \sqrt{MSE}$ and $t(\alpha/2m, N-k)$ is the upper $(\alpha/2m)$ th quantile of the t distribution with $N-k$ degrees of freedom. Bonferroni's simultaneous intervals guarantee that the probability of all the m confidence statements being correct is at least $1-\alpha$.

Consider a clinical trial on the efficacy of a newly developed drug for treating patients with migraine headaches (Wang and Chow, 1995). The clinical trial is a double-blind three-arm parallel randomized study comparing the newly developed drug (denoted by drug A), an active control agent (denoted by drug B), and a placebo control (denoted by P). The primary efficacy endpoint is the pain relief score. The pain relief scores are obtained 60 minutes after the administration of treatment based on a visual pain relief scale ranging from 0 (no relief from pain) to 10 (complete relief from pain). For the purpose of illustration, Table 8.4.2 gives the pain relief scores for each treatment. From Table 8.4.2, the sum

Table 8.4.2 Pain Relief Scores for Three Treatments

Treatment	Pain Relief (Y_{ij})	n_i	\bar{Y}_i
Placebo	0.0, 1.0	2	0.50
Drug A	3.1, 2.7, 3.8	3	3.20
Drug B	2.3, 3.5, 2.8, 2.5	4	2.78

Note: 0.0 means no relief from pain.

Source: Wang and Chow (1995).

of squares (SSE) due to error and the sum of squares (SSA) due to treatment can be obtained as

$$SSE = \sum_{i=1}^k \sum_{j=-1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 = 1.95,$$

and

$$SSA = \sum_{i=1}^k n_i (\bar{Y}_i - \bar{Y}_{..})^2 = 9.70.$$

Thus the F test is given by

$$\begin{aligned} F &= \frac{SSA/(k-1)}{SSE/\sum_{i=1}^k (n_i-1)} \\ &= \frac{9.70/2}{1.95/6} = 15.15, \end{aligned}$$

which is greater than $F(0.05, 2, 6) = 5.14$. Hence we can reject the null hypothesis of no treatment difference, $\mu_A = \mu_B = \mu_P$ at the $\alpha = 0.05$ level of significance. The analysis of variance table is summarized in Table 8.4.3.

Since the overall F test rejects the null hypothesis of no treatment differences, we can compare the test drug A , with the active control agent (i.e., drug B) and the test drug A with the placebo P by constructing Bonferroni's simultaneous confidence intervals for $\mu_A - \mu_B$ and $\mu_A - \mu_P$. The resulting confidence intervals are given by $(-1.32, 0.48)$ and $(1.13, 3.43)$, respectively (also see Table 8.4.4). Since the simultaneous confidence interval for $\mu_A - \mu_P$ does not contain 0, it indicates that there is significant difference between the test drug A and the placebo. On the other hand, there is no significant difference between the test drug A and the active control agent B because its simultaneous confidence interval contains 0.

Table 8.4.3 Analysis of Variance Table for Pain Relief Data

Source of Variation	df	Sum of Squares	Mean Sum of Squares	F
Treatment	2	9.70	4.85	15.15
Error	6	1.95	0.32	
Total	8	11.65		

Table 8.4.4 Bonferroni's 95% Simultaneous Confidence Intervals

Contrast	95% Simultaneous Confidence Interval
$\mu_A - \mu_B$	(-1.32, 0.48)
$\mu_A - \mu_P$	(1.13, 3.43)

Source: Wang and Chow (1995).

Two-Way Classification

As discussed in the previous section, clinical trials are often conducted at several sites. The purpose of a multicenter trial is not only to expedite the patient enrollment but also to confirm that the result is reproducible and hence can be generalized to the targeted patient population. However, if there is treatment-by-center interaction, the reproducibility and generalizability of the result are questionable. In practice, the following two-way classification model is considered to test the existence of the treatment-by-center interaction:

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + e_{ijk}, \quad (8.4.5)$$

where $i = 1, \dots, a$, $j = 1, \dots, b$, $k = 1, \dots, n$. In the above model τ_i is the true effect of the i th treatment, β_j is the true effect of the j th center, $(\tau\beta)_{ij}$ denotes the effect of the interaction between τ_i and β_j , and e_{ijk} is the random error in observing Y_{ijk} . For simplicity, both treatment and center effects are assumed fixed here. Since the treatment effects are defined as deviations from the overall mean, we have

$$\sum_{i=1}^a \tau_i = 0 \quad \text{and} \quad \sum_{j=1}^b \beta_j = 0.$$

Similarly, the interaction effects are fixed and defined so that

$$\sum_{i=1}^a (\tau\beta)_{ij} = \sum_{j=1}^b (\tau\beta)_{ij} = 0.$$

Define $\bar{Y}_{i..}$, $\bar{Y}_{j..}$, $\bar{Y}_{ij..}$, $\bar{Y}_{...}$ as the corresponding treatment, center, treatment-by-center, and overall sample mean. Then the deviation of an individual observation from the overall sample mean can be described as

$$\begin{aligned} Y_{ijk} - \bar{Y}_{...} &= (\bar{Y}_{i..} - \bar{Y}_{...}) + (\bar{Y}_{j..} - \bar{Y}_{...}) \\ &\quad + (\bar{Y}_{ij..} - \bar{Y}_{i..} - \bar{Y}_{j..} + \bar{Y}_{...}) + (Y_{ijk} - \bar{Y}_{ij..}). \end{aligned} \quad (8.4.6)$$

As a result, the deviation of an individual observation from the overall sample mean is partly due to differences among the means of the treatments, centers, treatment-by-center, and partly due to random variation in the measurements. Expression (8.4.6) leads to the following partitions of sum of squares (SS):

$$\begin{aligned} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{...})^2 &= bn \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2 + an \sum_{j=1}^b (\bar{Y}_{j..} - \bar{Y}_{...})^2 \\ &\quad + n \sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{ij..} - \bar{Y}_{i..} - \bar{Y}_{j..} + \bar{Y}_{...})^2 \\ &\quad + \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{ij..})^2. \end{aligned}$$

The first term on the left-hand side is the total sum of squares and the first two terms on the right-hand side are sums of squares due to the treatment and the center, denoted by SSA and SSC, respectively, while the last two terms are sums of squares due to treatment-by-center and random error, denoted by SS(AC) and SSE, respectively. Therefore, we have

$$\text{SST} = \text{SSA} + \text{SSC} + \text{SS(AC)} + \text{SSE}.$$

Assuming that treatment and center are fixed, the expected values of the mean squares are given by

$$E(\text{MSA}) = E\left(\frac{\text{SSA}}{a-1}\right) = \sigma^2 + \frac{bn\sum_{i=1}^a \tau_i^2}{a-1},$$

$$E(\text{MSC}) = E\left(\frac{\text{SSC}}{b-1}\right) = \sigma^2 + \frac{an\sum_{j=1}^b \beta_j^2}{b-1},$$

$$E[\text{MS(AC)}] = E\left(\frac{\text{SS(AC)}}{(a-1)(b-1)}\right) = \sigma^2 + \frac{n\sum_{i=1}^a \sum_{j=1}^b (\tau_i \beta_j)^2}{(a-1)(b-1)},$$

$$E(\text{MSE}) = E\left(\frac{\text{SSE}}{ab(n-1)}\right) = \sigma^2.$$

To test the hypotheses $H_0: \tau_i = 0$ (i.e., no treatment effect), $H_0: \beta_j = 0$ (i.e., no center effect), and $H_0: (\tau\beta)_{ij} = 0$ (i.e., no treatment-by-center interaction), we can divide the corresponding mean square by mean square error. This ratio will follow an F distribution with appropriate numerator degrees of freedom and $ab(n-1)$ denominator degrees of freedom. Table 8.4.5 summarizes the analysis of variance table for the two-way classification fixed model.

8.5 ANALYSIS OF COVARIANCE

In parallel-group clinical trials, patients who meet inclusion and exclusion criteria are randomly assigned to each treatment group. Under the assumption that the targeted patient population is homogeneous, we can expect that patient characteristics such as age, gender, and weight are comparable between treatment groups. If the patient population is known to be heterogeneous in terms of some demographic variables, then a stratified randomization according to these variables should be applied. At the beginning of the study, clinical data are usually collected at randomization to establish baseline values. After the administration of study drug, clinical data are often collected at each visit over the entire course of study. These clinical data are then analyzed to assess the efficacy and safety of the treatments.

As pointed out earlier, before the analysis of endpoint values, the comparability of patient characteristics between treatments is usually examined by an analysis of variance if the variable is continuous. For the analysis of endpoint values, although the technique of analysis of variance can be directly applied, it is believed the endpoint values are usually linearly related to the baseline values. Therefore an adjusted analysis of variance should be considered to account for the baseline values. This adjusted analysis of variance is called analysis of covariance (ANCOVA).

Table 8.4.5 Analysis of Variance for the Two-Way Classification Fixed Model

Source of Variation	Sum of Squares	df	Mean Squares	F
Treatment	$SSA = bn \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2$	$a - 1$	$MSA = \frac{SSA}{(a - 1)}$	$F_1 = \frac{MSA}{MSE}$
Center	$SSC = an \sum_{j=1}^b (\bar{Y}_{.j} - \bar{Y}_{...})^2$	$b - 1$	$MSC = \frac{SSC}{b - 1}$	$F_2 = \frac{MSC}{MSE}$
Treatment * center	$SS(AC) = n \sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j} + \bar{Y}_{...})^2$	$(a - 1)(b - 1)$	$MS(AC) = \frac{SS(AC)}{(a - 1)(b - 1)}$	$F_3 = \frac{MS(AC)}{MSE}$
Error	$SSE = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{ij.})^2$	$ab(n - 1)$	$MSE = \frac{SSE}{ab(n - 1)}$	
Total	$SST = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{...})^2$	$abn - 1$		

Let X_{ij} and Y_{ij} be the baseline and endpoint values for the j th patient in the i th treatment group, where $j = 1, \dots, n_i$ and $i = 1, \dots, k$. To account for the baseline values, model (8.4.2) can be modified as follows:

$$Y_{ij} = \mu + \tau_i + \beta X_{ij} + e_{ij}$$

or

$$Y_{ij} = \bar{Y}_{..} + \tau_i + \beta(X_{ij} - \bar{X}_{..}) + \varepsilon_{ij}. \quad (8.5.1)$$

Under the above model, the least squares estimators of μ , τ_i , and β are given by

$$\begin{aligned}\hat{\mu} &= \bar{Y}_{..}, \\ \hat{\tau}_i &= \bar{Y}_{i..} - \bar{Y}_{..} - \hat{\beta}(\bar{X}_{i..} - \bar{X}_{..}),\end{aligned}$$

and

$$\hat{\beta} = \frac{E_{XY}}{E_{XX}},$$

where $\bar{X}_{i..}$ and $\bar{Y}_{i..}$ denote the sample means of baseline and endpoint for the i th treatment, respectively, and

$$\begin{aligned}E_{XX} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i..})^2, \\ E_{XY} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i..})(Y_{ij} - \bar{Y}_{..}).\end{aligned}$$

To derive a test for the null hypothesis of no treatment difference as described in (8.4.1), we denote

$$\begin{aligned}S_{XX} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})^2, \\ S_{XY} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..}), \\ S_{YY} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2, \\ T_{XX} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{X}_{i..} - \bar{X}_{..})^2, \\ T_{XY} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{X}_{i..} - \bar{X}_{..})(\bar{Y}_{i..} - \bar{Y}_{..}), \\ T_{YY} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{Y}_{i..} - \bar{Y}_{..})^2.\end{aligned}$$

Table 8.5.1 Analysis of Covariance

Source of Variation	Sum of Squares	df	Mean Squares
Regression	$\frac{(E_{XY})^2}{E_{XX}}$	1	
Treatment	SSA	$k - 1$	$MSA = \frac{SSA}{k - 1}$
Error	SSE	$N - k - 1$	$MSE = \frac{SSE}{N - k - 1}$
Total	$SST = S_{YY}$	$N - 1$	

Note: See the text for definitions of sums of squares and cross products.

Note that it can be verified that

$$E_{XX} = S_{XX} - T_{XX},$$

$$E_{XY} = S_{XY} - T_{XY},$$

$$E_{YY} = S_{YY} - T_{YY}.$$

Therefore, the sum of squares due to treatment and error can be obtained as follows:

$$SSA = S_{YY} - \frac{S_{XY}^2}{S_{XX}} - SSE,$$

$$SSE = E_{YY} - \frac{E_{XY}^2}{E_{XX}}.$$

The analysis of covariance table for one-way classification is provided in Table 8.5.1. Under the null hypothesis of no treatment difference, namely $\tau_i = 0$ for all i , the test statistic

$$F = \frac{SSA/(k - 1)}{SSE/(N - k - 1)}$$

follows an F distribution with $(k - 1, N - k - 1)$ degrees of freedom, where $\sum_{i=1}^k n_i$. Therefore we can reject the null hypothesis of no treatment difference if

$$F > F(\alpha, k - 1, N - k - 1).$$

Note that if $\beta = 1$, then model (8.5.1) reduces to

$$Y_{ij} = \mu + \tau_i + X_{ij} + e_{ij}$$

or

$$Y_{ij} - X_{ij} = \mu + \tau_i + e_{ij}, \quad (8.5.2)$$

where $Y_{ij} - X_{ij}$ denotes the change from the baseline of endpoint for the j th patient in the i th treatment group. Therefore, we can perform the usual analysis of variance on $\{Y_{ij} - X_{ij}\}$. This analysis is known as a change from the baseline analysis of variance.

In practice, some researchers often include the change from the baseline analysis of covariance by treating the baseline as a covariate. This leads to the following model:

$$Y_{ij} - X_{ij} = \mu + \tau_i + \beta X_{ij} + e_{ij}$$

or

$$Y_{ij} - X_{ij} = \mu + \tau_i + \beta(X_{ij} - \bar{X}_{..}) + \varepsilon_{ij}. \quad (8.5.3)$$

It should be noted that the above model is equivalent to model (8.5.1), since it can be rewritten as

$$Y_{ij} = \mu + \tau_i + \beta^*(X_{ij} - \bar{X}_{..}) + \varepsilon_{ij}, \quad (8.5.4)$$

where $\beta^* - 1 = \beta$. In other words, under model (8.5.1), we are interested in testing the hypothesis that $H_0: \beta = 0$, which is equivalent to testing the hypothesis that $H_0: \beta^* = 1$ under model (8.5.4).

8.6 NONPARAMETRICS

In the previous section, the statistical inferences derived from the analysis of variance or the analysis of covariance are primarily based on the normality assumption of random errors. In practice, the normality assumption is usually not met. In this case the validity of the statistical inference drawn is questionable. As an alternative, we may consider using the method of nonparametrics to draw a statistical inference. In other words, no distribution assumptions are imposed. In this section, for illustration purposes, we will focus on the Wilcoxon signed rank test for one sample and rank sum test for shift in location due to treatment between two populations. This concept is also extended to derive the Kruskal-Wallis test for comparing k treatment groups. Distribution-free tests for other complicated design situations will not be covered in this section. However, the details of these tests can be found in Hollander and Wolfe (1973) and Lehmann (1975).

Wilcoxon Signed Rank Test

Let Y_{Bi} and Y_{Ei} be the response of the i th patient at baseline and endpoint, respectively, where $i = 1, \dots, n$. Suppose that one of the study objectives is to show that there is a significant difference between pre- and post-treatment in patient's response. A nonparametric approach for this purpose is to consider the model

$$D_i = \theta + e_i, \quad i = 1, \dots, n,$$

where $D_i = Y_{Ei} - Y_{Bi}$, θ is the unknown treatment effect, and e_i are random errors which are assumed to be mutually independent and symmetric about zero. To test the null hypothesis of no treatment effect,

$$H_0: \theta = 0,$$

we may consider the following distribution-free signed rank test. First, form the absolute differences $|D_i|$, $i = 1, \dots, n$, and let R_i denote the rank of $|D_i|$ in the joint ranking from the least to the greatest of $|D_1|, \dots, |D_n|$. Define

$$\psi_i = \begin{cases} 1 & \text{if } D_i > 0, \\ 0 & \text{if } D_i < 0. \end{cases}$$

We then reject the null hypothesis at the α level of significance if

$$T^+ \leq t(\alpha_2, n) \quad \text{or} \quad T^+ \geq \frac{n(n+1)}{2} - t(\alpha_1, n),$$

where

$$T^+ = \sum_{i=1}^n R_i \psi_i$$

and the constant $t(\alpha, n)$ satisfies the equation

$$P\{T^+ \leq t(\alpha, n)\} = \alpha.$$

Note that the values of $t(\alpha, n)$ are given in Appendix A.6. The above distribution-free signed rank test is usually referred to as a Wilcoxon signed rank test. Under the null hypothesis, the statistic

$$T^* = \frac{T^+ - E(T^+)}{\text{var}(T^+)}$$

has an asymptotic standard normal distribution where

$$E(T^+) = \frac{n(n+1)}{4}$$

and

$$\text{var}(T^+) = \frac{n(n+1)(2n+1)}{24}.$$

As a result, we may reject the null hypothesis for large samples if

$$|T^*| \geq Z(\alpha/2).$$

When there are ties, we can replace $\text{var}(T^+)$ with the following

$$\text{var}(T^+) = \frac{1}{24} \left[n(n+1)(2n+1) - \frac{1}{2} \sum_{j=1}^g t_j(t_j-1)(t_j+1) \right],$$

where g is the number of tied groups and t_j is the size of the tied group j .

Wilcoxon Rank Sum Test

A distribution-free approach can also be applied to test whether there is a significant difference between two treatments. Let $\{Y_{1i}, i = 1, \dots, m\}$ and $\{Y_{2i}, i = 1, \dots, n\}$ be two

random samples from population 1 and population 2, respectively. Consider the model

$$Y_{1i} = e_i, \quad i = 1, \dots, m,$$

and

$$Y_{2j} = e_{m+j} + \Delta, \quad j = 1, \dots, n,$$

where Δ denotes the unknown shift in location due to the treatment. In other words, the null hypothesis of no treatment difference can be formulated as

$$\begin{aligned} H_0: \quad & \Delta = 0, \\ \text{vs.} \quad & H_a: \quad \Delta \neq 0. \end{aligned}$$

To derive a distribution-free test for the above null hypothesis, we assume that $e_i, i = 1, \dots, n+m$ are identically distributed and are mutually independent. The Wilcoxon rank sum test can be summarized as follows: First, we order the $N = n+m$ observations from the smallest to the largest and let R_j denote the rank of Y_{2j} in this order. Then we let

$$R = \sum_{i=1}^n R_j.$$

then, the Wilcoxon–Mann–Whitney test statistic is given by

$$W = R - \frac{n(n+1)}{2}$$

Thus, the statistic W is the sum of the ranks assigned to the Y_{2j} 's centered by $n(n+1)/2$ which follows a distribution symmetric about $nm/2$. As a result it can be verified that

$$w(1 - \alpha, m, n) = nm - w(\alpha, m, n),$$

where the constant $w(\alpha, m, n)$ satisfies

$$P\{W \leq w(\alpha, m, n)\} = \alpha.$$

Note that values of $w(\alpha, m, n)$ are given in Appendix A.5 which gives the critical values in the Mann–Whitney form of Wilcoxon rank sum statistic, that is, $W - n(n+1)/2$.

Based on the statistic W , we can reject the null hypothesis of no treatment difference at the α level of significance if

$$W \leq w(\alpha_2, m, n) \quad \text{or} \quad W \geq [nm - w(\alpha_1, m, n)],$$

where $\alpha = \alpha_1 + \alpha_2$. For a one-sided test, we reject $H_a: \Delta \leq 0$ if

$$W \leq w(\alpha, m, n),$$

and reject $H_a: \Delta > 0$ if

$$W \geq nm - w(\alpha, m, n),$$

at the α level of significance. Note that when there are ties among the N observations, W must be calculated based on average ranks.

For large samples, by the Central Limit Theorem, the test statistic

$$W^* = \frac{W - E(W)}{[\text{var}(W)]^{1/2}}$$

has an asymptotic standard normal distribution as $\min(m, n)$ tends to infinity, where $E(W)$ and $\text{var}(W)$ are the expected value and variance of W , which are given by

$$E(W) = \frac{nm}{2},$$

$$\text{var}(W) = \frac{mn(m+n+1)}{12}.$$

Therefore, we can reject the null hypothesis of no shift in location due to treatment at the α level of significance if

$$|W^*| \geq Z(\alpha/2).$$

In the case where there are ties, we can replace $\text{var}(W)$ with

$$s^2 = \frac{mn}{12} \left[m + n + 1 - \frac{\sum_{j=1}^g t_j(t_j^2 - 1)}{(m+n)(m+n-1)} \right],$$

where g is the number of tied groups and t_j is the size of tied group j (Hollander and Wolfe, 1973). Note that an isolated observation is considered to be a tied group of size 1.

To illustrate the application of the Wilcoxon rank sum test, consider a clinical trial comparing the safety and efficacy of three drugs (A , B , and C) in the treatment of acute sinusitis in adults. This trial was a multicenter parallel-group randomized study. A ten-point assessment questionnaire (TAQ) is used to capture the improvement of signs and symptoms. The set of TAQ consists of ten questions. Each question is intended to capture the improvement of a specific symptom. A list of the ten symptoms is given in Table 8.6.1. For each question a score of 0 represents absence and a score 2 indicates presence. Table 8.6.2 lists partial data of the total TAQ scores over the ten questions at the baseline. Based on data given in Table 8.6.2, pairwise comparisons can be made using a Wilcoxon rank sum

Table 8.6.1 Ten Symptoms in TAQ

Item	Symptom
1	Fever
2	Nasal discharge
3	Nasal congestion
4	Cough
5	Headache
6	Facial pain
7	Facial swelling
8	Reduced activity
9	Impaired sleep
10	Impaired appetite

Table 8.6.2 TAQ Scores

Treatment	Subject Number	TAQ Score	Rank of TAQ Score
A	121	6	1.0
A	169	14	15.5
A	521	8	3.0
A	522	12	9.0
A	526	14	15.5
A	527	14	15.5
A	529	14	15.5
A	531	16	22.5
A	569	14	15.5
A	571	10	6.0
B	525	20	29.0
B	528	16	22.5
B	530	12	9.0
B	532	14	15.5
B	533	12	9.0
B	570	16	22.5
B	572	14	15.0
B	575	18	26.0
B	576	18	26.0
B	577	20	29.0
C	289	8	3.0
C	523	16	22.5
C	524	8	3.0
C	573	20	29.0
C	574	18	26.0
C	689	10	6.0
C	690	14	15.5
C	693	10	6.0
C	694	14	15.5
C	697	14	15.5

test. The results are summarized below:

$$A \text{ versus } B: W_1 = \sum_{j=1}^n R_{1j} - 55 = 134 - 55 = 79,$$

$$A \text{ versus } C: W_2 = \sum_{j=1}^n R_{2j} - 55 = 112 - 55 = 57,$$

$$B \text{ versus } C: W_3 = \sum_{j=1}^n R_{3j} - 55 = 85 - 55 = 30$$

where W_i denotes the sum of the ranks assigned to the Y 's for the i th comparison. Since $w(0.975, 10, 10) = (10)(10) - w(0.025, 10, 10) = 100 - 24 = 76$, W_1 is the only test that is greater than $w(0.975, 10, 10)$. Therefore, we reject the null hypothesis of no difference between treatments A and B . There are no significant differences between treatments A and C and between treatments B and C . Note that a normal approximation gives

$$A \text{ versus } B: s_1 = 76, W^* = -2.21, p = 0.027,$$

$$A \text{ versus } C: s_2 = 98, W^* = -0.51, p = 0.610,$$

$$B \text{ versus } C: s_3 = 125, W^* = 1.49, p = 0.135.$$

Hence, the normal approximation also indicates that there is a significant difference between treatments A and B .

Kruskal-Wallis Test

When there are more than two treatments to compare, a similar idea can be carried out to derive a test for the following hypothesis

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k.$$

Let Y_{ij} be the observation for the j th patient in the i th treatment group. Consider model (8.4.2) repeated below:

$$Y_{ij} = \mu + \tau_i + e_{ij}, \quad j = 1, \dots, n_i, i = 1, \dots, k.$$

We first rank all $N = \sum_{i=1}^k n_i$ observations jointly, from the least to the greatest. Let r_{ij} be the rank of Y_{ij} in this joint ranking. Also, for $i = 1, \dots, k$, denote

$$R_i = \sum_{j=1}^{n_i} r_{ij}, \quad \bar{R}_i = \frac{R_i}{n_i}, \quad \text{and} \quad \bar{R} = \frac{N+1}{2}.$$

then the test statistic

$$\begin{aligned} H &= \frac{12}{N(N+1)} \sum_{i=1}^k n_i (\bar{R}_i - \bar{R})^2 \\ &= \left(\frac{12}{N(N+1)} \sum_{i=1}^k \frac{\bar{R}_i^2}{n_i} \right) - 3(N+1) \end{aligned}$$

can be used to test the null hypothesis of no treatment differences. We reject the null hypothesis at the α level of significance if

$$H \geq h(\alpha, k, (n_1, n_2, \dots, n_k)),$$

where the constant $h(\alpha, k, (n_1, n_2, \dots, n_k))$ satisfies

$$P\{H \geq h(\alpha, k, (n_1, n_2, \dots, n_k))\} = \alpha.$$

The values of $h(\alpha, k, (n_1, n_2, \dots, n_k))$ can be found in Iman, Quade, and Alexander (1975). Note that when there are ties, H should be calculated based on the average ranks.

When $\min\{n_1, \dots, n_k\}$ tends to infinity, H is approximately distributed as a χ^2 distribution with $k-1$ degrees of freedom. Therefore, for large samples we reject the null hypothesis of no treatment differences if

$$H \geq \chi^2(\alpha, k-1),$$

where $\chi^2(\alpha, k-1)$ is the upper α percentile of a χ^2 distribution with $k-1$ degrees of freedom. When there are ties, we replace H with H^* as follows:

$$H^* = \left\{ 1 - \left(\frac{\sum_{i=1}^g (t_i^3 - t_i)}{N^3 - N} \right) \right\}^{-1} H,$$

where t_i and g are defined as before.

To illustrate the application of Kruskal-Wallis test for comparing k treatments, consider the example concerning total TAQ scores described earlier. From 8.6.2, the Kruskal-Wallis test statistic can be easily obtained as

$$H = \left(\frac{12}{N(N+1)} \sum_{i=1}^k \frac{\bar{R}_i^2}{n_i} \right) - 3(N+1) = 5.215,$$

which is approximately χ^2 distributed with 2 degrees of freedom. Therefore we fail to reject the null hypothesis of no treatment differences at the 5% level of significance level since the observed p -value is given by 0.074.

8.7 REPEATED MEASURES

In clinical trials multiple assessments of a response variable are often performed at various time points (visits) from each of the patients under study. As a result the collected clinical data set may consist of repeated observations and a set of covariates (e.g., time and some patient characteristics) for each of the patients. This type of data is usually referred to as *longitudinal* data.

In many clinical trials, repeated observations after the administration of drug products are necessarily obtained to assess the efficacy and safety of the drug products under study. The objectives of repeated measures are (1) to determine whether the optimal therapeutic effect has been reached, (2) to determine whether a dose titration is necessary for good clinical practice, (3) to monitor the progress and/or health-related quality of life of patients with chronic diseases such as cancer, or (4) to study the behavior of the study drug over time (or to detect whether a potential pattern or trend in time exists).

In practice, different statistical models can be applied to address the above objectives under different model assumptions. In this section we will introduce some commonly used statistical models for the analysis of repeated measures in clinical trials. These models include the usual analysis of variance models for assessment of overall average drug effect across time points, for detection of time (visit) effect, and for determination of treatment-by-time effect and the method of generalized estimating equations (GEE) proposed by Zeger and Liang (1986) and Liang and Zeger (1986).

Assessment of Overall Average Effect Across Time

In clinical trials, the ultimate goal is to determine whether there is a drug effect in terms of efficacy and safety of the drug product under study. Repeated measures occur in each patient. For simplicity, we may consider the *visit* as nested within the *patient*. Then a nested model can be used to assess the overall average drug effect across time points (visits). Let Y_{ijk} be the observation from the k th visit for the j th patient who is in the i th treatment group, where $i = 1, \dots, a$, $j = 1, \dots, n$, and $k = 1, \dots, m$. The nested model used to describe Y_{ijk} is

$$Y_{ijk} = \mu + \tau_i + P_{j(i)} + e_{ijk}, \quad (8.7.1)$$

where μ is the overall mean, τ_i and $P_{j(i)}$ denote the fixed effect of the i th treatment and the effect due to the j th patient within the i th treatment, respectively, and e_{ijk} is the random

error in observing Y_{ijk} . It is assumed that e_{ijk} are independent and identically normally distributed with mean 0 and variance σ^2 . Note that in the complete randomized design, $P_{j(i)}$ is often expressed as

$$P_{j(i)} = P_j + (AP)_{ij},$$

where P_j and $(AP)_{ij}$ denote the effect due to the j th patient and the effect of the interaction between the j th patient and the i th treatment.

Similarly under model (8.7.1) the deviation of an individual observation from the overall sample mean can be partitioned as follows:

$$Y_{ijk} - \bar{Y}_{...} = (\bar{Y}_{i..} - \bar{Y}_{...}) + (\bar{Y}_{ij.} - \bar{Y}_{i..}) + (Y_{ijk} - \bar{Y}_{ij.})$$

Therefore we have

$$\text{SST} = \text{SSA} + \text{SSP}(A) + \text{SSE},$$

where SSA, SSP(A), and SSE denote sum of squares due to treatment, patient within treatment, and error, respectively, with associated degrees of freedoms given as

$$(ann - 1) = (a - 1) + a(n - 1) + an(m - 1).$$

As usual, dividing the sum of squares of error by its corresponding degrees of freedom gives the mean squares. The analysis of variance table of model (8.7.1) is given in Table 8.7.1.

It can be seen from Table 8.7.1 that under the normality assumption, the test statistic

$$F_A = \frac{\text{SSA}/(a - 1)}{\text{SSE}/[an(m - 1)]}$$

is distributed as an F distribution with $(a - 1, an(m - 1))$ degrees of freedom. Therefore, we reject the null hypothesis of no treatment difference; that is, $H_0: \tau_i = 0$, for all i if

$$F_A \geq F[\alpha, a - 1, an(m - 1)].$$

Similarly, the effect due to patient within treatment can also be tested by

$$F_{P(A)} = \frac{\text{SSP}(A)/[a(n - 1)]}{\text{SSE}/[an(m - 1)]}$$

which has an F distribution with $[a(n - 1), an(m - 1)]$ degrees of freedom. Note that if there is a significant difference effect due to the patients within treatment, it indicates that the responses to treatment are not consistent from patient to patient. To further investigate the response of each patient, one might examine the patient means within each treatment based on the Newmann-Keuls range test (Keuls, 1952).

Detection of Time Effect

It is clear that model (8.7.1) does not provide any information regarding the effect due to time or visit. To examine whether there is a significant effect due to visit, we may further

Table 8.7.1 Analysis of Variance for Model (8.7.1)

Source of Variation	Sum of Squares	df	Mean Square
Treatment	$SSA = mn \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2$	$a - 1$	$MSA = \frac{SSA}{(a - 1)}$
Patient (Treatment)	$SSP(A) = m \sum_{i=1}^a \sum_{j=1}^n (\bar{Y}_{ij} - \bar{Y}_{i..})^2$	$a(n - 1)$	$MSP(A) = \frac{SSP(A)}{a(n - 1)}$
Error	$SSE = \sum_{i=1}^a \sum_{j=1}^n \sum_{k=1}^m (Y_{ijk} - \bar{Y}_{ij.})^2$	$an(m - 1)$	$MSE = \frac{SSE}{an(m - 1)}$
Total	$SST = \sum_{i=1}^a \sum_{j=1}^n \sum_{k=1}^m (Y_{ijk} - \bar{Y}_{...})^2$	$anm - 1$	

partition the residual sum of squares from model (8.7.1) into sum of squares due to visit and sum of squares due to error. Therefore, model (8.7.1) can be modified as

$$Y_{ijk} = \mu + \tau_i + P_{j(i)} + V_k + \varepsilon_{ijk}, \quad (8.7.2)$$

where μ , τ_i , and $P_{j(i)}$ are as defined in model (8.7.1), V_k denotes the effect due to the k th visit (or assessment) and $\varepsilon_{ijk} = e_{ijk} - V_k$. Similarly, we can partition the deviation of an individual observation from the overall sample mean as

$$\begin{aligned} Y_{ijk} - \bar{Y}_{...} &= (\bar{Y}_{i..} - \bar{Y}_{...}) + (\bar{Y}_{ij.} - \bar{Y}_{i..}) + (\bar{Y}_{..k} - \bar{Y}_{...}) \\ &\quad + (Y_{ijk} - \bar{Y}_{ij.} - \bar{Y}_{..k} + \bar{Y}_{...}). \end{aligned}$$

Therefore, we have

$$\text{SST} = \text{SSA} + \text{SSP}(A) + \text{SSV} + \text{SSE},$$

where SSV denotes the sum of squares due to visit. The associated degrees of freedoms are

$$(anm - 1) = (a - 1) + a(n - 1) + (m - 1) + (an - 1)(m - 1).$$

As usual, dividing each independent sum of squares by its corresponding degrees of freedom gives the mean squares. The analysis of variance table of model (8.7.2) is given in Table 8.7.2.

As can be seen from Table 8.7.2, under the normality assumption, the test statistic

$$F_V = \frac{\text{SSV}/(m - 1)}{\text{SSE}/(an - 1)(m - 1)}$$

is distributed as an F distribution with $[m - 1, (an - 1)(m - 1)]$ degrees of freedom. Therefore, we reject the following null hypothesis:

$$H_0: v_1 = v_2 = \dots = v_m,$$

where v_k , $k = 1, \dots, m$, denote the mean of the k th visit.

If there is a time effect, we can use appropriate contrasts to test whether there is a linear or quadratic trend over time. Alternatively, we can consider *time* as a covariate and fit a least squares slope through each patient's data points. Then we can perform the usual analysis of variance based on these slopes by treating the slope as an outcome measure. Note that the analysis based on slopes for longitudinal lung function data is satisfactory.

Treatment-by-Time Interaction

In model (8.7.2) we consider the *visit* (or *time*) as a class variable and perform an analysis of variance to test to see whether there is a time effect. For simplicity, we assume that the effect due to the patient is fixed and that there is no treatment-by-time interaction. In many clinical trials, the responses of the patients are expected to vary from one to another. To account for between-patient variability and to test for possible treatment-by-time interaction, the following model is useful:

$$Y_{ijk} = \mu + \tau_i + P_{j(i)} + V_k + (\tau V)_{ik} + \varepsilon_{ijk}, \quad (8.7.3)$$

Table 8.7.2 Analysis of Variance for Model (8.7.2)

Source of Variation	Sum of Squares	df	Mean Squares
Treatment	$SSA = mn \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2$	$a - 1$	$MSA = \frac{SSA}{(a - 1)}$
Patient (Treatment)	$SSP(A) = m \sum_{i=1}^a \sum_{j=1}^n (\bar{Y}_{ij.} - \bar{Y}_{i..})^2$	$a(n - 1)$	$MSP(A) = \frac{SSP(A)}{a(n - 1)}$
Visit	$SSV = an \sum_{k=1}^m (\bar{Y}_{.k} - \bar{Y}_{...})^2$	$m - 1$	$MSV = \frac{SSV}{m - 1}$
Error	$SSE = \sum_{i=1}^a \sum_{j=1}^n \sum_{k=1}^m (Y_{ijk} - \bar{Y}_{ij.} - \bar{Y}_{.k} - \bar{Y}_{...})^2$	$(an - 1)(m - 1)$	$MSE = \frac{SSE}{(an - 1)(m - 1)}$
Total	$SST = \sum_{i=1}^a \sum_{j=1}^n \sum_{k=1}^m (Y_{ijk} - \bar{Y}_{...})^2$	$ann - 1$	

where μ , τ_i , V_k , and ε_{ijk} are as defined in model (8.7.2), $P_{j(i)}$ is the random effect of the j th patient within the i th treatment, and $(\tau V)_{ik}$ denotes the effect due to the interaction between the i th treatment and the k th visit (or assessment). It is assumed that $P_{j(i)}$ are i.i.d. normal with mean 0 and variance σ_S^2 , which are mutually independent of ε_{ijk} . Similarly, we can partition the deviation of an individual observation from the overall sample mean as follows:

$$\begin{aligned} Y_{ijk} - \bar{Y}_{...} &= (\bar{Y}_{i..} - \bar{Y}_{...}) + (\bar{Y}_{ij.} - \bar{Y}_{i..}) + (\bar{Y}_{..k} - \bar{Y}_{...}) \\ &\quad + (\bar{Y}_{i,k} - \bar{Y}_{i..} - \bar{Y}_{..k} + \bar{Y}_{...}) + (Y_{ijk} - \bar{Y}_{ij.} - \bar{Y}_{i,k} + \bar{Y}_{i..}). \end{aligned}$$

Therefore, we have

$$\text{SST} + \text{SSA} + \text{SSP}(A) + \text{SSV} + \text{SS}(AV) + \text{SSE},$$

where $\text{SS}(AV)$ denotes the sum of squares due to the interaction between treatment and visit. The associated degrees of freedoms are

$$(anm - 1) = (a - 1) + a(n - 1) + (m - 1) + (a - 1)(m - 1) + a(n - 1)(m - 1).$$

The analysis of variance table of model (8.7.3) is given in Table 8.7.3. To test the hypotheses of interest, we can divide the corresponding mean square by mean square error. This ratio will follow an F distribution with appropriate numerator degrees of freedom and denominator degrees of freedom. For example, we can reject $H_0: \tau_i = 0$ for all i (i.e., no treatment effect) if

$$F_A = \frac{\text{SSA}/(a - 1)}{\text{SSP}(A)/[a(n - 1)]} \geq F[\alpha, a - 1, a(n - 1)],$$

where $F[\alpha, a - 1, a(n - 1)]$ is the upper α th quantile of an F distribution with $[(a - 1), a(n - 1)]$ degrees of freedom. Similarly, we would reject $H_0: (\tau V)_{ik} = 0$ (i.e., no treatment-by-visit interaction) if

$$\begin{aligned} F_{AV} &= \frac{\text{SS}(AV)/[(n - 1)(m - 1)]}{\text{SSE}/[a(n - 1)(m - 1)]} \\ &\geq F[\alpha, (n - 1)(m - 1), a(n - 1)(m - 1)]. \end{aligned}$$

Note that an overall assessment of the average drug effect across visits is possible if the treatment-by-visit interaction is not qualitatively significant. When an interaction between treatment and visit is observed, it indicates that the differences in treatment are not consistent across visits. This inconsistency may be caused by (1) the fact that the time to reach optimal therapeutic effect varies from one drug to another, (2) the dose titration procedure (if any) is not done adequately, or (3) the disease status of the patients change over time. In this case it is suggested that the time effect or pattern be carefully evaluated in order to make valid statistical inference regarding the efficacy of the drug under study.

Method of Generalized Estimating Equations (GEE)

Thus far we have discussed several linear models for assessing (1) overall average drug effect across time points, (2) the detection of time effect, and (3) the effect due to treatment-by-time interaction in clinical trials with repeated measures. The primary assumptions of these linear

Table 8.7.3 Analysis of Variance for Model (8.7.3)

Source of Variation	Sum of Squares	df	Mean Squares
Treatment	$SSA = mn \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2$	$a - 1$	$MSA = \frac{SSA}{(a - 1)}$
Patient (Treatment)	$SSP(A) = m \sum_{i=1}^a \sum_{j=1}^n (\bar{Y}_{ij.} - \bar{Y}_{i..})^2$	$a(n - 1)$	$MSP(A) = \frac{SSP(A)}{a(n - 1)}$
Visit	$SSV = an \sum_{k=1}^m (\bar{Y}_{.k} - \bar{Y}_{...})^2$	$m - 1$	$MSV = \frac{SSV}{(m - 1)}$
Treatment* Visit	$SS(AV) = \sum_{i=1}^a \sum_{j=1}^n \sum_{k=1}^m (\bar{Y}_{ijk} - \bar{Y}_{i..} - \bar{Y}_{.k} + \bar{Y}_{...})^2$	$(a - 1)(m - 1)$	$MS(AV) = \frac{SS(AV)}{(a - 1)(m - 1)}$
Error	$SSE = \sum_{i=1}^a \sum_{j=1}^n \sum_{k=1}^m (Y_{ijk} - \bar{Y}_{ij.} - \bar{Y}_{i..} + \bar{Y}_{...})^2$	$a(n - 1)(m - 1)$	$MSE = \frac{SSE}{a(n - 1)(m - 1)}$
Total	$SST = \sum_{i=1}^a \sum_{j=1}^n \sum_{k=1}^m (Y_{ijk} - \bar{Y}_{...})^2$	$ann - 1$	

models are (1) the outcome follows a normal distribution, (2) the outcome does not vary across time and/or across subjects, and (3) repeated observations are independent. Note that with independent observations, generalized linear models (GLMs) can be extended for time-dependent data in a variety of ways (McCullagh and Nelder, 1983). In addition, the parameters in a GLM can also be assumed to vary across time as a stochastic process and/or across subjects according to a mixing distribution (Zeger and Qaqish, 1988). More recently Vonesh and Chinchilli (1997) provide a comprehensive review of the various methods for analyzing repeated measurements.

In clinical trials, however, the outcome may not follow a normal distribution and may vary across time and/or across subjects. In addition, repeated outcomes for one individual are usually correlated with one another. Hence, to obtain a valid statistical inference, the analysis of longitudinal data must take into account for the correlation among repeated observations within each subject. Failure to account for the correlation may result in inefficient estimates of the parameters and/or inconsistent estimates with respect to precision across subjects. As a result, the issue of correlation among repeated observations for a subject provides a challenge for analysis of longitudinal data.

To account for the correlation among a subject's repeated observations, Zeger and Liang (1992) and Diggle, Liang, and Zeger (1994) point out that the following three basic models are useful: marginal, transition, and random effects models. For a marginal model, the regression coefficients for the clinical response on covariates and the structure of the intra-subject correlation are modeled separately. The marginal expectation which is referred to as the average response over a particular stratum with the common values of covariates being a function of the explanatory variables. The regression coefficients in the marginal model hence have the same interpretation as those obtained from a cross-sectional analysis. The variance in the marginal model is the product of a variance function of the marginal mean and a scale parameter. The intrasubject correlation at two different time points is also a function of marginal means at the two time points and possible additional parameters. The approach of transition models tries to model both regression and intrasubject correlation simultaneously. Hence, a common equation includes, on the same scale, both unknown parameters for the dependence of the clinical endpoints on explanatory variables and for the intrasubject correlation. The transition model then assumes that the conditioned expectation at a particular time point is a function of covariates and the proceeding responses. Consequently the prior responses are treated as explicit predictors the same way as any other explanatory variables. The concept of a random effects model is to consider subjects enrolled in the study as a representative random sample from the targeted population. As a result, the regression coefficients of a subject in the trial is also a random vector that is assumed to follow a probability distribution. Ware et al. (1988) discuss the conceptual and technical differences between the marginal and transition models.

Zeger and Liang (1986) and Liang and Zeger (1986) propose a method using either the marginal or transition model in conjunction with the quasi-likelihood approach (McMullagh and Nelder, 1983; Wedderburn, 1974) in order to obtain consistent estimates of the parameters and their corresponding variances under rather weak assumptions on the correlation among a subject's repeated observations. This method is known as the generalized estimating equations (GEE). Under some assumptions for the first two moments of the joint distribution of correlated clinical responses, the regression coefficients, including the treatment effects and the correlation parameter, can be estimated consistently by solving a multivariate analogue of the quasi-score function without invoking an intractable likelihood with many nuisance parameters.

The quasi-likelihood approach is a regression methodology that requires few assumptions about the distribution of the dependent variable, and hence it can be used with a variety of outcomes. The quasi-likelihood approach is usually applied to the marginal model, which describes the relationship between the outcome and a set of explanatory variables, in order to obtain estimating equations of the parameters of interest. Based on the estimating equations, consistent estimates of the parameters and their corresponding variances can be obtained.

Let Y_{it} be the observation of the i th subject observed at time t , where $t = 1, \dots, n_i$. Also, let \mathbf{X}_{it} be a $p \times 1$ vector of explanatory variables (or covariates) related to Y_{it} . These explanatory variables could be patient characteristics such as age and gender or time (visit). The marginal model relates the marginal expectation of Y_{it} with respect to \mathbf{X}_{it} by

$$g(\mu_{it}) = \mathbf{X}'_{it} \boldsymbol{\beta},$$

where g is a known link function, $\mu_{it} = E(Y_{it})$, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters. It is also assumed that the marginal variance of Y_{it} is a function of the marginal mean:

$$\text{var}(Y_{it}) = v(\mu_{it})\phi,$$

where v is a known function and ϕ is the overdispersion parameter that accounts for the variation of Y_{it} not explained by $v(\mu_{it})$. In addition, the covariance between Y_{is} and Y_{it} , $s < t = 1, \dots, n_i$, is assumed to be a function of the marginal means and additional parameter $\boldsymbol{\eta}$:

$$\text{cov}(Y_{is}, Y_{it}) = c(\mu_{is}, \mu_{it}, \boldsymbol{\eta}),$$

where c is a known function and $\boldsymbol{\eta}$ is a $s \times 1$ vector of unknown parameters that measures the within-subject correlation. Note that the estimate of $\boldsymbol{\beta}$ is often used to describe how the averaged response (rather than one subject's response) changes with respect to the covariates. Based on the above assumptions, the consistent estimator (or quasi-likelihood estimator) is the solution of the following scorelike estimating equation system.

$$\sum_{i=1}^K \mathbf{D}'_i \mathbf{V}_i^{-1} \mathbf{S}_i = 0,$$

where

$$\begin{aligned} \mathbf{S}_i &= \mathbf{Y}_i - \boldsymbol{\mu}_i \\ \mathbf{D}_i &= \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \end{aligned}$$

with $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})'$ and $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{in_i})'$ and

$$\mathbf{V}_i = \frac{\mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\eta}) \mathbf{A}_i^{1/2}}{\phi},$$

in which \mathbf{A}_i is an $n_i \times n_i$ diagonal matrix with $g(\mu_{ij})$ as the j th diagonal element and $\mathbf{R}_i(\boldsymbol{\eta})$ is the $n_i \times n_i$ working correlation matrix for each \mathbf{Y}_i . Note that we do not expect $\mathbf{R}_i(\boldsymbol{\eta})$ to be correctly specified. The GEE method provides consistent estimates even when $\mathbf{R}_i(\boldsymbol{\eta})$ is incorrect.

Zeger and Liang (1992) indicate that the GEE method enjoys at least three useful properties. First of all, the estimated regression coefficients are nearly as efficient as the maximum likelihood estimates for most clinical trials. Second, despite a possible misspecification of the covariance structure for the correlated clinical responses, the estimates for regression coefficients are still consistent if the sample size is sufficiently large. Finally, the statistical inference of the regression coefficients will generally not be influenced by the covariance matrix of the clinical responses as long as the robust estimated covariance matrix of the regression coefficient estimates suggested by Liang and Zeger (1986) is used.

Example 8.7.1 For illustration purposes, consider a clinical trial that compares the effect of gender on the distance (in millimeters) from the center of the pituitary to the pterygomaxillary fissure. Table 8.7.4 lists measurements obtained from 20 children (10 girls and 10 boys) at ages of 8, 10, 12, and 14. Figure 8.7.1 plots the mean distances at ages of 8, 10, 12, and 14; and the plot suggests that there is a significant difference between boys and girls. To assess the overall average effect, age effect, and gender-by-age interaction, we consider models (8.7.1)–(8.7.3). The analyses of variance results are summarized in Table 8.7.5. From Table 8.7.3 it can be seen that there is a marginally significant gender-by-age interaction (p -value = 0.06). This quantitative interaction, however, does not involve the analysis by pooling data across ages (see also Figure 8.7.1). Model (8.7.1) provides an overall assessment of the gender effect. The result indicates that there is a significant difference between boys and girls (p -value < 0.01). In addition model (8.7.2) reveals that there is an age effect. To further compare the age effect between boys and girls, we can try to fit the linear regression for all boys and all girls and then perform an analysis of variance on the estimated slopes.

Table 8.7.4 Measurements on 10 Girls and 10 Boys at 4 Different Ages

Gender	Subject	Age in Years			
		8	10	12	14
Girl	1	22.130	21.100	22.645	24.190
	2	22.130	22.645	25.220	26.765
	3	21.615	25.220	25.735	27.280
	4	24.705	25.735	26.250	27.795
	5	22.645	24.190	23.675	24.705
	6	21.100	22.130	22.130	23.675
	7	22.645	23.675	24.190	26.250
	8	24.190	24.190	24.705	25.220
	9	21.100	22.130	23.160	22.645
	10	17.495	20.070	20.070	20.585
Boy	1	27.280	26.250	30.370	32.430
	2	22.645	23.675	24.190	27.795
	3	24.190	23.675	25.220	28.825
	4	26.765	28.825	27.795	28.310
	5	21.100	24.705	23.675	27.280
	6	25.735	26.765	28.310	29.855
	7	23.160	23.160	25.735	27.795
	8	25.220	22.645	25.735	26.765
	9	24.190	21.615	32.430	27.280
	10	28.825	29.340	32.430	32.945

Table 8.7.5 Analyses of Variance for Data in Table 8.7.4

Model	Objective	Source of Variation	df	Sum of Squares	F	p-Value
8.7.1	Overall average effect	Gender	1	202.26	48.05	<0.01
		Subject (gender)	18	297.48	3.93	<0.01
8.7.2	Age effect	Gender	1	202.26	110.35	<0.01
		Subject (gender)	18	297.48	9.02	<0.01
		Age	3	148.09	26.93	<0.01
8.7.3	Gender-by-age effect	Gender	1	202.26	119.76	<0.01
		Subject (gender)	18	297.48	9.79	<0.01
		Age	3	148.09	29.23	<0.01
		Gender-by-age	3	13.27	2.62	0.06

As described in this section, the major disadvantage of the analysis of variance models (8.7.1)–(8.7.3) is that they assume that the measurements within each subject are independent. To account for the correlation among measurements within each subject, the method of GEE is also considered. To enhance efficiency of the estimate, the following working correlation structure helps describe the correlation between observations within a subject

$$\begin{bmatrix} 1.00 & 0.687 & 0.687 & 0.687 \\ 0.687 & 1.000 & 0.687 & 0.687 \\ 0.687 & 0.687 & 1.000 & 0.687 \\ 0.687 & 0.687 & 0.687 & 1.000 \end{bmatrix}.$$

The estimates, standard errors, and Z-scores obtained by the method of GEE are summarized in Table 8.7.6. In Table 8.7.6 all of the parameters are statistically significant

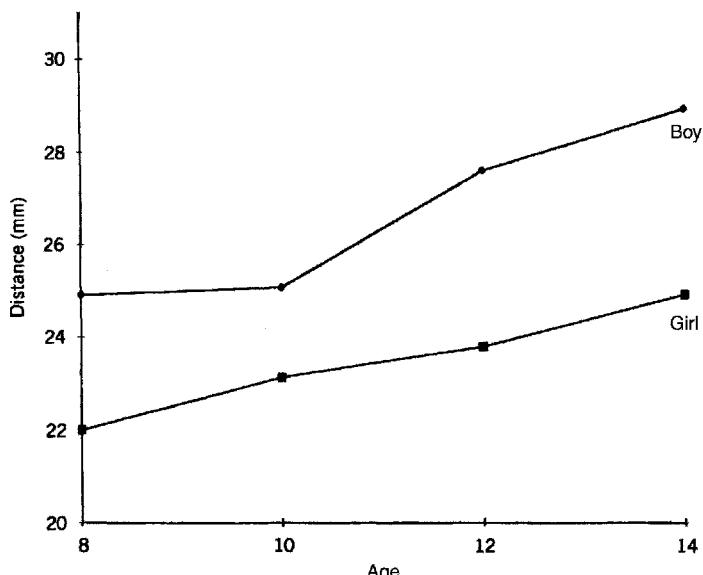
**Figure 8.7.1** Mean distances from the pituitary to the pterygomaxillary fissure in boys and girls.

Table 8.7.6 Estimates, Standard Errors, and Z-Scores from the Method of GEE

Parameter	Estimate	se Naive	se Robust	Z Robust
Intercept	28.928	0.697	0.651	44.46
Gender	-4.017	0.986	0.936	-4.29
Age				
A_1	-4.017	0.551	0.459	-8.74
A_2	-3.863	0.551	0.570	-6.78
A_3	-1.339	0.551	0.775	-1.73
Gender*Age				
$(GA)_1$	1.082	0.780	0.626	1.73
$(GA)_2$	2.060	0.780	0.677	3.04
$(GA)_3$	0.206	0.780	0.808	0.26

except for A_3 (with Z-score = 1.73), $(GA)_1$ (with Z-score = 1.73), and $(GA)_3$ (with Z-score = 0.26).

8.8 DISCUSSION

As was indicated earlier, a two-sample t test is useful in comparing two treatment means and an F test can be applied to compare k treatments. There is relationship between the t test statistic and the F test statistic when $k = 2$. That is, $t^2 = F$. In other words, the F test reduces to the two-sample t test when there are only two groups evaluated in a clinical trial.

For nonparametric analysis the Wilcoxon rank sum test and the Kruskal-Wallis test are fairly robust against outliers because they are based on ranks (or relative order) rather than absolute difference of the raw data. Both Wilcoxon rank sum test and Kruskal-Wallis test can also be obtained by replacing the raw data with their corresponding ranks in the t test and sum of squares for treatment (i.e., SSA) in a one-way analysis of variance.

For the GEE method, the principal distinction between the marginal model and random effects model is whether the regression coefficients describe an individual's change or an average response change with respect to the covariates. As a result, Zeger, Liang, and Albert (1988) classify these two approaches as population-average (marginal) and subject-specific (random effects) models. The population-average approach describes how the average response across subjects changes with covariates, while the subject-specific method accounts for the heterogeneity among subjects to estimate subject-specific coefficients. In addition to the GEE method for the analysis of longitudinal data in clinical trials, there are others discussed in the literature. For example, Grizzle and Allen (1969) and Hui (1984) consider fitting growth curves to repeated observations for each subject. When there are missing and/or unequally spaced examinations, Rosner and Munoz (1988) suggest that an autoregressive model be used.

The traditional approach to repeated measurements in clinical trials is to perform an analysis of variance at the last visit. This method is usually referred to as endpoint analysis or last observation carried forward (LOCF) analysis. The result is compared with those obtained by the analyses of variance performed at each visit. If the results from the analysis at each visit are consistent with the endpoint analysis, then an overall conclusion on the treatment effect is drawn. However, if there is a significant difference among visits and/or the last visit, then a further analysis is performed to determine whether there is a pattern

and/or trend over time. Note that a discussion regarding the validity of the LOCF analysis in clinical trials can be found in Cheng and Chow (2003).

A commonly encountered problem in clinical trials with repeated measures is missing data. In general, missing data may have two causes. On one hand, subjects may withdraw prematurely from the study at a post-treatment time point. These subjects are referred to as dropouts. As a result the data scheduled to be collected since the last visit would be completely missing. On the other hand, subjects might be lost to follow-up at a scheduled visit. However, they may return either on another unscheduled date or on the next scheduled visit. These subjects may or may not complete the study. Diggle et al. (1994) refer to those missing values due to the missed visits as intermittently missing values. Several methods have been proposed that assume the missing values to occur at random within treatment groups. See, for example, Crépeau et al. (1985), Diggle (1988, 1989), Kenward (1987), and Ridout (1991). In clinical trials it is suggested that the dropout rates and missing patterns across all visits be compared between treatments. If there is a discrepancy in dropout rates or a missing pattern, it is crucial to investigate whether a response-dependence or treatment-related problem has occurred (see also Chow and Shao, 2002a).

9

ANALYSIS OF CATEGORICAL DATA

9.1 INTRODUCTION

For the assessment of the effectiveness and safety of an investigational pharmaceutical entity, ideally the scale for the primary clinical endpoint should be numerically continuous to provide an accurate and reliable assessment. In practice, however, it is impossible, or can be extremely expensive, to measure responses quantitatively. On the other hand, patients' responses to treatments can be easily documented according to the occurrence of some meaningful and well-defined event such as death, infection, or cure of a certain disease and any serious adverse events. In addition the intensity of these events can be graded according to some predefined categories. Therefore categorical data can be useful surrogate endpoints for some unobserved latent continuous variables in clinical trials. Sometimes, to provide an easy analysis and/or a better presentation of the results, continuous data are transformed to categorical data with respect to some predefined criteria. As a result many efficacy and safety endpoints in clinical trials are in the form of categorical data on either a nominal or ordinal scale. In this section we introduce the concept and methodology for analysis of categorical data through some real examples from published results of clinical trials. The first example concerns the U.S. Physicians' Health Study on the prevention of cardiovascular mortality and fatal and nonfatal myocardial infarction with a low dose of aspirin (325 mg on alternate days). Another primary goal of the U.S. Physicians' Study was to investigate the impact of long-term supplementation with beta carotene (50 mg on alternate day) on the incidence of malignant neoplasm and cardiovascular disease. Table 9.1.1 shows the number of subjects of malignant neoplasms for the treatment assignment of the beta carotene component (Hennekens et al., 1996). As can be seen in Table 9.1.1, the subjects were classified into two categories according to the occurrence of malignant neoplasm over an average of 12 years of

Table 9.1.1 Number of Subjects With Malignant Neoplasms in the Beta Carotene Component of the U.S. Physician's Health Study

Malignant Neoplasm	Beta Carotene	Placebo
<i>N</i>	11,036	11,035
Yes	1273	1293
Year 1–2	120	130
Year 3–4	157	136
Year 5–9	500	567
≥10 years	496	460
No	9763	9742

Source: Hennekens et al. (1996).

treatment and follow-up. A clinical endpoint of this kind is referred to as a *binary response* on a nominal scale with two categories.

The next example is from a parallel-group, randomized, double-blind, placebo-controlled trial conducted by the U.S. National Institute of Neurological Disorders and Stroke (NINDS) to investigate the efficacy and safety of intravenous recombinant tissue plasminogen activator (rt-PA) for ischemic stroke. This study consisted of two parts. The first part was to examine the clinical activity of rt-PA in terms of early clinical improvement as compared to the placebo. For the second part, the primary hypothesis of interest was if there was a consistent and persuasive difference between the rt-PA and placebo groups in terms of the proportion of patients who recovered with minimal or no deficit three months after treatment. In addition to the study center, stratified randomization was performed according to time from the onset of stroke to the start of treatment (0 to 90 or 91 to 180 minutes). One of the primary clinical efficacy endpoints was early clinical improvement, which is defined as at least a four-point improvement from the baseline values in National Institutes of Health Stroke Scale (NIHSS) score or a resolution of the neurological deficit 24 hours after the onset of stroke. Hence, the subjects of this study are also classified into two groups based on the defined clinical endpoint, namely those with the defined clinical improvement and those without. The defined clinical improvement is a binary response. However, unlike the occurrence of malignant neoplasm in the U.S. Physicians' Study, it is a composite clinical index derived from changes of other endpoints over the entire course of treatment. Table 9.1.2 relates the number of subjects with clinical improvement by study part and time.

In order to provide an objective, scientific, and clinically meaningful assessment of efficacy in the therapy for benign prostatic hyperplasia (BPH), the American Urological Association developed and validated a symptom index for BPH which is the sum of scores over seven individual symptoms. This symptom index is usually referred to as the AUA-7 symptom score. Table 9.1.3 lists the seven questionnaires associated with these symptom indexes. Each question in the AUA-7 symptom score consists of six categories, which are arranged in a monotone increasing order. This type of categorical data is said to be in the ordinal scale and is sometimes referred to as *ordered categorical data*. The results of the initial question 4 on frequency in the validation of discrimination between the BPH patients from the controls are displayed in Table 9.1.4. Another example of the use of the ordered categorical data is to evaluate efficacy and safety of a newly developed contrast agent in conjunction with magnetic resonance imaging for diagnosis of malignant liver lesions. The

Table 9.1.2 Number of Subjects Showing Clinical Improvement

	rt-PA	Placebo
<i>Part 1 0–90 min</i>		
<i>N</i>	71	68
Yes	36	31
No	35	37
<i>91–180 min</i>		
<i>N</i>	73	79
Yes	31	26
No	42	53
<i>Part 2 0–90 min</i>		
<i>N</i>	86	77
Yes	51	30
No	35	47
<i>90–180 min</i>		
<i>N</i>	82	88
Yes	29	35
No	53	53

Source: National Institute of Neurological Disorders and rt-PA Stroke Study Group (1995).

primary diagnostic efficacy endpoint is the change in diagnostic confidence compared to precontrast MRI scan. This endpoint consists of a total of six categories on the ordinal scale: worsened, unchanged, minimal improvement, moderate improvement, good improvement, and excellent improvement. Table 9.1.5 provides the summary of change in diagnostic confidence by treatment group and gender from a hypothetical dataset.

The time-response profile is critical in the assessment of a new proposed therapy for the treatment of a certain illness. Evaluations of effectiveness and safety for each subject participating in the trial are performed at a number of preselected time points during entire course of the study. As a result, the categorical data collected from most clinical trials are always *repeated categorical endpoints* that are not statistically independent within the same subjects. For some known or unknown reasons, subjects might not come to the clinic for a particular scheduled visit or subjects might simply drop out and never return to the study. Therefore, repeated categorical data with missing data is the norm for most clinical trials. Koch et al., (1990) report partial results from a randomized, multicenter, placebo-controlled trial in the assessment of a new test drug for a respiratory disorder. The raw data provided by Koch et al. (1990) is an ordinal categorical endpoint of each subject's status measured at baseline and at four prescheduled visits after the initiation of treatment. This ordinal index, referred as the status score in this chapter, consists of five categories: terrible, poor, fair, good, and excellent. A summary of the transition from baseline to visit 3 in terms of frequency is provided in Table 9.1.6.

Table 9.1.3 AUA Symptom Index

Questions	Not at All	Less than 1 Time in 5	Less than Half the Time	About Half the Time	More than Half the Time	Almost Always
1. During the last month or so, how often have you a sensation of not emptying your bladder completely after you finished urinating?	0	1	2	3	4	5
2. During the last month or so, how often have you had to urinate again less than 2 hours after you finished urinating?	0	1	2	3	4	5
3. During the last month or so, how often have you stopped and started again several times when you urinated?	0	1	2	3	4	5
4. During the last month or so, how often have you found it difficult to postpone urination?	0	1	2	3	4	5
5. During the last month or so, how often have you had a weak urinary stream?	0	1	2	3	4	5
6. During the last month or so, how often have you had to push or strain to begin urination?	0	1	2	3	4	5
Questions	None	1 Time	2 Times	3 Times	4 Times	5 or More Times
7. During the last month, how many times did you most typically get up to urinate from the time you went to bed at night until the time you got up in the morning?	0	1	2	3	4	5

AUA-7 Symptom score is the sum of scores of questions 1 to 7.

Source: Barry et al. (1992).

Table 9.1.4 Partial Results of Validation for Question 4—Frequency 1

Point Score	BPH	Controls
<i>N</i>	73	56
0	8	16
1	18	23
2	17	7
3	14	7
4	8	2
5	8	1

Source: Barry et al. (1992).

In general, different statistical methods are required for the analysis of different categorical data types. This chapter will not only cover descriptive statistics but also interval estimation and hypotheses testing for one-sample, two-sample, and multiple-sample categorical data. These statistics are derived based on random assignment of patients to the treatment. Since, as discussed in Chapter 4, most randomization-based methods currently available are for hypothesis testing and the scope of inference is limited only to the subjects in the study, they must be adequately supplemented with a model-based method for estimation. The advantages of model-based methods include (1) the ability for investigation of the homogeneity of treatment effects across strata, (2) the flexibility for adjustment of demographic and baseline characteristics, and (3) the reduction of variability (or the increase of power) by inclusion of covariates associated with the categorical endpoints. The major shortcomings of the model-based methods consists in the inability to prove the underlying assumption required for inference and a greater difficulty in the implementation and interpretation of the results. As a result, it may be a good idea to combine both the randomization-based method and the model-based method for analysis of categorical data.

In the next section, statistical analyses of binary data for one sample and comparison within paired samples are introduced. Statistical inference based on binary data for independent samples are given in Section 9.3. Section 9.4 outlines statistical methods for ordered categorical data. Section 9.5 provides the methodology for combining results from the analysis of categorical data over strata. We note that most methods covered in Section 9.2

Table 9.1.5 Summary by Treatment for Change in Diagnostic Confidence

	Placebo		Test Drug	
	Female	Male	Female	Male
<i>N</i>	37	36	36	37
Worsened	4	4	1	1
Unchanged	11	9	1	2
Improvement				
Minimal	7	8	2	1
Moderate	4	6	2	3
Good	6	4	12	13
Excellent	5	5	18	17

Table 9.1.6 Summary of Frequencies of Transition of Status Score from the Baseline to Visit 3

Baseline	Terrible	Poor	Fair	Good	Excellent	Total
<i>Treatment: Placebo</i>						
Terrible	0	0	0	0	0	0
Poor	4	2	3	1	1	11
Fair	4	1	9	4	2	20
Good	1	3	2	4	9	19
Excellent	2	0	0	2	3	7
Total	11	6	14	11	15	57
<i>Treatment: Test Drug</i>						
Terrible	0	0	2	1	0	3
Poor	0	3	2	1	4	10
Fair	2	0	5	3	8	18
Good	0	0	0	7	5	12
Excellent	0	0	1	2	8	11
Total	2	3	10	14	25	54

through Section 9.5 are randomization-based methods. Model-based methods such as the logistic regression model are discussed in Section 9.6. In Section 9.7 we review some methods for the analysis of repeated categorical data. Some final remarks and a brief discussion are given in Section 9.8.

9.2 STATISTICAL INFERENCE FOR ONE SAMPLE

As described in the previous section, for the NINDS trial, there are two possible outcomes for each subject based on either the predefined criteria of at least a four-point improvement from the baseline in NIHSS score or a complete resolution of the neurological deficit. The two possible outcomes are either a subject improved or did not improve. For this study one of the primary objectives is to evaluate whether subjects treated with rt-PA have a greater proportion of early improvement compared to the placebo group. Before a formal statistical comparison between the rt-PA and placebo is made, a descriptive statistic for estimation of the unknown proportion of the subjects who improved in each group must be obtained.

Let Y_i denote the clinical endpoint that indicates whether a subject had an improvement in a group of n subjects treated with rt-PA, namely

$$Y_i = \begin{cases} 1, & \text{if subject had an improvement,} \\ 0, & \text{otherwise.} \end{cases} \quad (9.2.1)$$

where $i = 1, \dots, n$. Since the n subjects in rt-PA group are different and are not related to each other, the occurrence of improvement of a subject is random and is independent of that of other subjects. In other words, Y_i are statistically independent for $i = 1, \dots, n$. Under the assumption of independence, the population proportion of the subjects with improvement, denoted by P , can be estimated by the sample proportion of the subjects

with improvement, denoted by p , which is calculated as the number of subjects with improvement divided by the total number of subjects treated with rt-PA. In other words,

$$\begin{aligned} p &= \frac{1}{n} \{ \text{number of subjects with improvement} \} \\ &= \frac{1}{n} \sum_{i=1}^n Y_i \\ &= \frac{Y}{n}, \end{aligned} \quad (9.2.2)$$

where

$$Y = \sum_{i=1}^n Y_i$$

represents the number of subjects (out of n subjects) showing improvement who were treated with rt-PA. Y is usually referred to as a binomial random variable. Since Y_i has only two possible values, 0 and 1, the sample mean $p = Y/n$ is the sample proportion of subjects with improvement. The variance of p can be estimated by

$$v(p) = \frac{1}{n} p(1 - p). \quad (9.2.3)$$

It follows that the statistic

$$Z = \frac{p - P}{\sqrt{v(p)}} \quad (9.2.4)$$

has approximately a standard normal distribution when the sample size is large, for example, greater than 30. Based on Z , a $(1-\alpha) \times 100\%$ confidence interval for the population proportion can be obtained as

$$p \pm Z(\alpha/2) \sqrt{v(p)}, \quad (9.2.5)$$

where $Z(\alpha/2)$ is the upper $(\alpha/2)$ th quantile of the standard normal distribution.

Note that the confidence interval given in (9.2.4) will not give an adequate coverage probability when the sample size is small or the number of subjects with improvement is close to 0 or to n . When n is small, Clopper and Pearson (1934) suggest using the so-called *exact confidence interval* to find the binomial proportion. This method, however, is quite conservative in the sense that the actual coverage probability is greater than the nominal level of $(1-\alpha)100\%$. As a result, this method can yield a relatively wide confidence interval. Alternatively, Blyth and Still (1983) propose a procedure for the binomial proportion. Their proposed interval provides an adequate coverage probability. In addition, the proposed interval has approximately equal tail probabilities. Table 9.2.1 reproduces the lower and upper limits of the 95% confidence intervals for n from 1 to 30. When n is greater than 30, they suggest using the formula

$$\begin{aligned} &\frac{1}{[n + Z^2(\alpha/2)]} \\ &\times \left\{ (Y \pm 0.5) + Z^2(\alpha/2) \pm Z(\alpha/2) \sqrt{(Y \pm 0.5) - \frac{1}{n} (Y \pm 0.5)^2 + \frac{1}{4} Z^2(\alpha/2)} \right\}. \end{aligned} \quad (9.2.6)$$

For $Y = 0$ or 1, the lower limits are given as either 0 or $1 - (1 - \alpha)^{1/n}$ with an equivalent exception for $Y = n, n - 1$.

Table 9.2.1 Lower and Upper 95% Confidence Intervals for the Binomial Proportion

Sample Size	Total Responses	Lower Limit	Upper Limit	Sample Size	Total Responses	Lower Limit	Upper Limit
1	0	0.00	0.95	9	4	0.17	0.75
1	1	0.05	1.00	9	5	0.25	0.83
2	0	0.00	0.78	9	6	0.32	0.90
2	1	0.03	0.97	9	7	0.44	0.96
2	2	0.22	1.00	9	8	0.56	0.99
3	0	0.00	0.63	9	9	0.68	1.00
3	1	0.02	0.86	10	0	0.00	0.29
3	2	0.14	0.98	10	1	0.01	0.44
3	3	0.37	1.00	10	2	0.04	0.56
4	0	0.00	0.53	10	3	0.09	0.62
4	1	0.01	0.75	10	4	0.15	0.70
4	2	0.10	0.90	10	5	0.22	0.78
4	3	0.25	0.99	10	6	0.29	0.85
4	4	0.47	1.00	10	7	0.38	0.91
5	0	0.00	0.50	10	8	0.44	0.96
5	1	0.01	0.66	10	9	0.56	0.99
5	2	0.08	0.81	10	10	0.71	1.00
5	3	0.19	0.92	11	0	0.00	0.26
5	4	0.34	0.99	11	1	0.00	0.40
5	5	0.50	1.00	11	2	0.03	0.50
6	0	0.00	0.41	11	3	0.08	0.60
6	1	0.01	0.59	11	4	0.14	0.67
6	2	0.06	0.73	11	5	0.20	0.74
6	3	0.15	0.85	11	6	0.26	0.80
6	4	0.27	0.94	11	7	0.33	0.86
6	5	0.41	0.99	11	8	0.40	0.92
6	6	0.59	1.00	11	9	0.50	0.97
7	0	0.00	0.38	11	10	0.60	1.00
7	1	0.01	0.55	11	11	0.74	1.00
7	2	0.05	0.66	12	0	0.00	0.25
7	3	0.13	0.77	12	1	0.00	0.37
7	4	0.23	0.87	12	2	0.03	0.46
7	5	0.34	0.95	12	3	0.07	0.54
7	6	0.45	0.99	12	4	0.12	0.63
7	7	0.62	1.00	12	5	0.18	0.71
8	0	0.00	0.36	12	6	0.24	1.76
8	1	0.01	0.50	12	7	0.29	0.82
8	2	0.05	0.64	12	8	0.37	0.88
8	3	0.11	0.71	12	9	0.46	0.93
8	4	0.19	0.81	12	10	0.54	0.97
8	5	0.29	0.89	12	11	0.63	1.00
8	6	0.36	0.95	12	12	0.75	1.00
8	7	0.50	0.99	13	0	0.00	0.23
8	8	0.64	1.00	13	1	0.00	0.34
9	0	0.00	0.32	13	2	0.03	0.43
9	1	0.01	0.44	13	3	0.07	0.52
9	2	0.04	0.56	13	4	0.11	0.59
9	3	0.10	0.68	13	5	0.17	0.66

Table 9.2.1 (*Continued*)

Sample Size	Total Responses	Lower Limit	Upper Limit	Sample Size	Total Responses	Lower Limit	Upper Limit
13	6	0.22	0.74	16	9	0.30	0.80
13	7	0.26	0.78	16	10	0.37	0.82
13	8	0.34	0.83	16	11	0.44	0.87
13	9	0.41	0.89	16	12	0.50	0.91
13	10	0.48	0.93	16	13	0.56	0.95
13	11	0.57	0.97	16	14	0.63	0.98
13	12	0.66	1.00	16	15	0.70	1.00
13	13	0.77	1.00	16	16	0.80	1.00
14	0	0.00	0.23	17	0	0.00	0.19
14	1	0.00	0.32	17	1	0.00	0.28
14	2	0.03	0.42	17	2	0.02	0.35
14	3	0.06	0.50	17	3	0.05	0.42
14	4	0.10	0.58	17	4	0.08	0.49
14	5	0.15	0.63	17	5	0.12	0.54
14	6	0.21	0.68	17	6	0.17	0.59
14	7	0.24	0.76	17	7	0.19	0.65
14	8	0.32	0.79	17	8	0.25	0.72
14	9	0.37	0.85	17	9	0.28	0.75
14	10	0.42	0.90	17	10	0.35	0.81
14	11	0.50	0.94	17	11	0.41	0.83
14	12	0.58	0.97	17	12	0.46	0.88
14	13	0.68	1.00	17	13	0.51	0.92
14	14	0.77	1.00	17	14	0.58	0.95
15	0	0.00	0.22	17	15	0.65	0.98
15	1	0.00	0.30	17	16	0.72	1.00
15	2	0.02	0.39	17	17	0.81	1.00
15	3	0.06	0.47	18	0	0.00	0.18
15	4	0.10	0.53	18	1	0.00	0.27
15	5	0.14	0.61	18	2	0.02	0.33
15	6	0.19	0.67	18	3	0.05	0.41
15	7	0.22	0.71	18	4	0.08	0.47
15	8	0.29	0.78	18	5	0.12	0.53
15	9	0.33	0.81	18	6	0.16	0.59
15	10	0.39	0.86	18	7	0.18	0.63
15	11	0.47	0.90	18	8	0.24	0.67
15	12	0.53	0.94	18	9	0.27	0.73
15	13	0.61	0.98	18	10	0.33	0.76
15	14	0.70	1.00	18	11	0.37	0.82
15	15	0.78	1.00	18	12	0.41	0.84
16	0	0.00	0.20	18	13	0.47	0.88
16	1	0.00	0.30	18	14	0.53	0.92
16	2	0.02	0.37	18	15	0.59	0.95
16	3	0.05	0.44	18	16	0.67	0.98
16	4	0.09	0.60	18	17	0.73	1.00
16	5	0.13	0.56	18	18	0.82	1.00
16	6	0.18	0.63	19	0	0.00	0.17
16	7	0.20	0.70	19	1	0.00	0.25
16	8	0.27	0.73	19	2	0.02	0.32

Table 9.2.1 (*Continued*)

Sample Size	Total Responses	Lower Limit	Upper Limit	Sample Size	Total Responses	Lower Limit	Upper Limit
19	3	0.04	0.39	21	10	0.28	0.70
19	4	0.08	0.45	21	11	0.30	0.72
19	5	0.11	0.50	21	12	0.35	0.77
19	6	0.15	0.55	21	13	0.40	0.80
19	7	0.17	0.61	21	14	0.45	0.85
19	8	0.22	0.66	21	15	0.49	0.87
19	9	0.25	0.69	21	16	0.54	0.90
19	10	0.31	0.75	21	17	0.60	0.93
19	11	0.34	0.78	21	18	0.65	0.96
19	12	0.39	0.83	21	19	0.70	0.98
19	13	0.45	0.85	21	20	0.77	1.00
19	14	0.50	0.89	21	21	0.85	1.00
19	15	0.55	0.92	22	0	0.00	0.15
19	16	0.61	0.96	22	1	0.00	0.22
19	17	0.68	0.98	22	2	0.02	0.29
19	18	0.75	1.00	22	3	0.04	0.34
19	19	0.83	1.00	22	4	0.06	0.39
20	0	0.00	0.16	22	5	0.09	0.45
20	1	0.00	0.24	22	6	0.13	0.50
20	2	0.02	0.32	22	7	0.15	0.55
20	3	0.04	0.37	22	8	0.19	0.58
20	4	0.07	0.42	22	9	0.22	0.62
20	5	0.10	0.47	22	10	0.26	0.66
20	6	0.14	0.53	22	11	0.29	0.71
20	7	0.16	0.58	22	12	0.34	0.74
20	8	0.21	0.63	22	13	0.38	0.78
20	9	0.24	0.68	22	14	0.42	0.81
20	10	0.29	0.71	22	15	0.45	0.85
20	11	0.32	0.76	22	16	0.50	0.87
20	12	0.37	0.79	22	17	0.55	0.91
20	13	0.42	0.84	22	18	0.61	0.94
20	14	0.47	0.86	22	19	0.66	0.96
20	15	0.53	0.90	22	20	0.71	0.98
20	16	0.58	0.93	22	21	0.78	1.00
20	17	0.63	0.96	22	22	0.85	1.00
20	18	0.68	0.98	23	0	0.00	0.14
20	19	0.76	1.00	23	1	0.00	0.21
20	20	0.84	1.00	23	2	0.02	0.27
21	0	0.00	0.15	23	3	0.04	0.32
21	1	0.00	0.23	23	4	0.06	0.39
21	2	0.02	0.30	23	5	0.09	0.43
21	3	0.04	0.35	23	6	0.12	0.48
21	4	0.07	0.40	23	7	0.14	0.52
21	5	0.10	0.46	23	8	0.18	0.57
21	6	0.13	0.51	23	9	0.21	0.61
21	7	0.15	0.55	23	10	0.25	0.64
21	8	0.20	0.60	23	11	0.27	0.68
21	9	0.23	0.65	23	12	0.32	0.73

Table 9.2.1 (Continued)

Sample Size	Total Responses	Lower Limit	Upper Limit	Sample Size	Total Responses	Lower Limit	Upper Limit
23	13	0.36	0.75	25	12	0.30	0.68
23	14	0.39	0.79	25	13	0.32	0.70
23	15	0.43	0.82	25	14	0.36	0.75
23	16	0.48	0.86	25	15	0.40	0.78
23	17	0.52	0.88	25	16	0.44	0.81
23	18	0.57	0.91	25	17	0.48	0.84
23	19	0.61	0.94	25	18	0.52	0.87
23	20	0.68	0.96	25	19	0.56	0.89
23	21	0.73	0.98	25	20	0.60	0.92
23	22	0.79	1.00	25	21	0.64	0.94
23	23	0.86	1.00	25	22	0.70	0.97
24	0	0.00	0.13	25	23	0.75	0.99
24	1	0.00	0.20	25	24	0.81	1.00
24	2	0.02	0.26	25	25	0.87	1.00
24	3	0.03	0.31	26	0	0.00	0.12
24	4	0.06	0.37	26	1	0.00	0.19
24	5	0.09	0.41	26	2	0.01	0.24
24	6	0.11	0.46	26	3	0.03	0.20
24	7	0.13	0.50	26	4	0.05	0.34
24	8	0.17	0.54	26	5	0.08	0.38
24	9	0.20	0.59	26	6	0.11	0.42
24	10	0.23	0.63	26	7	0.12	0.47
24	11	0.26	0.66	26	8	0.15	0.51
24	12	0.31	0.69	26	9	0.19	0.54
24	13	0.34	0.74	26	10	0.21	0.58
24	14	0.37	0.77	26	11	0.24	0.62
24	15	0.41	0.80	26	12	0.28	0.66
24	16	0.46	0.83	26	13	0.30	0.70
24	17	0.50	0.87	26	14	0.34	0.72
24	18	0.54	0.89	26	15	0.38	0.76
24	19	0.59	0.91	26	16	0.42	0.79
24	20	0.63	0.94	26	17	0.46	0.81
24	21	0.69	0.97	26	18	0.49	0.85
24	22	0.74	0.98	26	19	0.53	0.88
24	23	0.80	1.00	26	20	0.58	0.89
24	24	0.87	1.00	26	21	0.62	0.92
25	0	0.00	0.13	26	22	0.66	0.95
25	1	0.00	0.19	26	23	0.80	0.97
25	2	0.01	0.25	26	24	0.76	0.99
25	3	0.03	0.30	26	25	0.81	1.00
25	4	0.06	0.36	26	26	0.88	1.00
25	5	0.08	0.40	27	0	0.00	0.12
25	6	0.11	0.44	27	1	0.00	0.18
25	7	0.13	0.48	27	2	0.01	0.23
25	8	0.16	0.52	27	3	0.03	0.29
25	9	0.19	0.56	27	4	0.05	0.33
25	10	0.22	0.60	27	5	0.08	0.37
25	11	0.25	0.64	27	6	0.10	0.41

Table 9.2.1 (*Continued*)

Sample Size	Total Responses	Lower Limit	Upper Limit	Sample Size	Total Responses	Lower Limit	Upper Limit
27	7	0.12	0.46	28	27	0.83	1.00
27	8	0.15	0.50	28	28	0.88	1.00
27	9	0.18	0.54	29	0	0.00	0.11
27	10	0.20	0.57	29	1	0.00	0.17
27	11	0.23	0.60	29	2	0.01	0.22
27	12	0.27	0.63	29	3	0.03	0.27
27	13	0.29	0.67	29	4	0.05	0.31
27	14	0.33	0.71	29	5	0.07	0.36
27	15	0.37	0.73	29	6	0.09	0.39
27	16	0.40	0.77	29	7	0.11	0.43
27	17	0.43	0.80	29	8	0.14	0.46
27	18	0.46	0.82	29	9	0.17	0.50
27	19	0.50	0.85	29	10	0.18	0.54
27	20	0.54	0.88	29	11	0.22	0.57
27	21	0.59	0.90	29	12	0.25	0.61
27	22	0.63	0.92	29	13	0.27	0.64
27	23	0.67	0.95	29	14	0.31	0.66
27	24	0.71	0.97	29	15	0.34	0.69
27	25	0.77	0.99	29	16	0.36	0.73
27	26	0.82	1.00	29	17	0.39	0.75
27	27	0.88	1.00	29	18	0.43	0.78
28	0	0.00	0.12	29	19	0.46	0.82
28	1	0.00	0.17	29	20	0.50	0.83
28	2	0.01	0.23	29	21	0.54	0.86
28	3	0.03	0.28	29	22	0.57	0.89
28	4	0.05	0.32	29	23	0.61	0.91
28	5	0.07	0.36	29	24	0.64	0.93
28	6	0.10	0.41	29	25	0.69	0.95
28	7	0.12	0.44	29	26	0.73	0.97
28	8	0.14	0.48	29	27	0.78	0.99
28	9	0.17	0.52	29	28	0.83	1.00
28	10	0.19	0.56	29	29	0.89	1.00
28	11	0.23	0.59	30	0	0.00	0.11
28	12	0.26	0.62	30	1	0.00	0.16
28	13	0.28	0.65	30	2	0.01	0.21
28	14	0.32	0.68	30	3	0.03	0.26
28	15	0.35	0.72	30	4	0.05	0.30
28	16	0.38	0.74	30	5	0.07	0.35
28	17	0.41	0.77	30	6	0.09	0.38
28	18	0.44	0.81	30	7	0.11	0.41
28	19	0.48	0.83	30	8	0.13	0.45
28	20	0.52	0.86	30	9	0.16	0.48
28	21	0.56	0.88	30	10	0.18	0.52
28	22	0.59	0.90	30	11	0.21	0.55
28	23	0.64	0.93	30	12	0.24	0.59
28	24	0.68	0.95	30	13	0.26	0.62
28	25	0.72	0.97	30	14	0.30	0.65
28	26	0.77	0.99	30	15	0.32	0.68

Table 9.2.1 (Continued)

Sample Size	Total Responses	Lower Limit	Upper Limit	Sample Size	Total Responses	Lower Limit	Upper Limit
30	16	0.35	0.70	30	24	0.62	0.91
30	17	0.38	0.74	30	25	0.65	0.93
30	18	0.41	0.76	30	26	0.70	0.95
30	19	0.45	0.79	30	27	0.74	0.97
30	20	0.58	0.82	30	28	0.79	0.99
30	21	0.52	0.84	30	29	0.84	1.00
30	22	0.55	0.87	30	30	0.89	1.00
30	23	0.59	0.89				

Source: Blyth and Still (1983).

Suppose that from the previous reports, about 30% to 35% of the subjects with acute ischemic stroke who are not treated with rt-PA improve. Hence, clinically it may be of interest to know whether the proportion of subjects treated with rt-PA is 35%. The following hypotheses can be used to address this question:

$$\begin{aligned} H_0: P = P_0 &= 0.35, \\ \text{vs. } H_a: P &\neq 0.35. \end{aligned} \quad (9.2.7)$$

Under the null hypothesis, $P = 0.35$, the large-sample test statistic is given by

$$Z = \frac{p - 0.35}{\sqrt{p(1-p)/n}}. \quad (9.2.8)$$

At a preselected level of significance α , we reject the null hypothesis and conclude that the proportion of subjects with improvement in those treated with rt-PA is different from 35% if

$$|Z| > Z(\alpha/2).$$

The corresponding p -value can be computed based on the standard normal distribution as follows:

$$p\text{-value} = P\{|Z| > z\}, \quad (9.2.9)$$

where z is the observed value of Z as given in (9.2.8). Therefore, the null hypothesis is rejected at the α level of significance if the p -value in (9.2.9) is smaller than α . Note that there is a relationship between hypothesis testing procedure given in (9.2.8) and the $(1-\alpha)100\%$ confidence interval given in (9.2.5). That is, one rejects the null hypothesis if and only if the $(1-\alpha)100\%$ confidence interval does not contain $P = 0.35$.

For small samples the exact test procedure based on p -value can be constructed. Under the null hypothesis Y follows a binomial distribution with $P = P_0 = 0.35$. We would expect an average of nP subjects treated with rt-PA to show improvement. If the observed Y is smaller than the expected number of subjects with improvement, the p -value can be calculated as

$$p\text{-value} = 2 \left\{ \sum_{y=0}^Y \frac{n!}{[y!(n-y)!]} (P_0)^y (1-P_0)^{n-y} \right\}. \quad (9.2.10)$$

On the other hand, when the observed Y is larger than the expected number of subject with improvement, nP , the p -value is given by

$$p\text{-value} = 2 \left\{ 1 - \sum_{y=0}^{Y-1} \frac{n!}{[y!(n-y)!]} (P_0)^y (1 - P_0)^{n-y} \right\}. \quad (9.2.11)$$

Similarly the null hypothesis is rejected at the α significance level if the p -value (calculated based on (9.2.10) or (9.2.11)) is smaller than α .

Example 9.2.1 Consider the dataset of the NINDS trial given in Table 9.2.1. The number of subjects with early improvement for rt-PA and the placebo group are 147 and 122, respectively. Since 312 subjects were enrolled in each group, the sample proportions with early improvement for rt-PA and placebo are 47.1% (147/312) and 39.1% (122/312), respectively. As presented in Table 9.2.2, the standard errors of estimated sample proportions are 0.0283 ($\sqrt{0.471 * 0.529/312}$) and 0.0276 ($\sqrt{0.391 * 0.609/312}$), respectively. The large-sample 95% confidence intervals for populations with early improvement for rt-PA and the placebo group are (41.58%, 52.65%) and (33.68%, 44.52%), respectively. The Blyth and Still method yields intervals of (41.49%, 52.82%) and (33.69%, 44.78%), which are 2.3% and 2.4% wider than those large-sample confidence intervals. The observed Z values for testing whether the proportion of improvement is 35% for rt-PA and the placebo are 4.287 ($((0.471 - 0.35)/0.0283)$) and 1.485 ($((0.391 - 0.35)/0.0276)$), respectively. As a result, since the observed Z value for the rt-PA group is greater than $z(0.025) = 1.96$ with the associated p -value of 0.00002, we reject the null hypothesis at the 5% level of significance and conclude that the proportion of subjects with early improvement for the rt-PA group is statistically significant greater than 35%. One can reach the same conclusion by noting that the 95% confidence interval (41.58%, 52.65%) does not contain 35%. For placebo group, since the observed Z value of 1.485 is less than 1.96, we fail to reject the null hypothesis. As a result, we conclude that the data of the placebo group do not provide sufficient evidence to doubt the validity of the null hypothesis that the proportion of the subjects with early improvement is equal to 35%. The same conclusion was reached based on (1) the corresponding p -value being 0.138, which is greater than 0.05, and (2) the 95%

Table 9.2.2 Summary of the Number of Subjects with Improvement by Treatment for the NINDS Trial

	rt-PA	Placebo
N	312	312
With improvement	147	122
Proportion	47.12%	39.10%
Standard error	2.83%	2.76%
95% Confidence Interval		
Large-sample	(41.58%, 52.65%)	(33.68%, 44.52%)
Blyth-Still	(41.49%, 52.82%)	(33.69%, 44.78%)
Z-statistics	4.287	1.485
p -value		
Large-sample	0.00002	0.1375
Exact	0.00001	0.1500

confidence interval for the placebo group of (33.68%, 44.52%) including $P = 35\%$. The exact p -value for the placebo group is given by

$$\begin{aligned} p\text{-value} &= 2 \left\{ 1 - \sum_{y=0}^{121} \frac{312!}{[y!(312-y)!]} (0.35)^y (0.65)^{312-y} \right\} \\ &= 0.150. \end{aligned}$$

Similarly, the exact p -value of rt-PA group is given by 0.00001. Note that in this example, both the exact and large-sample methods yield similar p -values. These results are summarized in Table 9.2.2.

In clinical trials, the comparison between the pre-treatment (baseline) and the post-treatment is usually made for the assessment of efficacy and safety of the treatment. This comparison within the same subjects is usually referred to as the pre-post comparison which is also known as *change from baseline analysis*. To illustrate the pre-post comparison concept, consider regrouping the five classes of the status score of the Koch's data set given in Table 9.1.6 into two categories: unfavorable (i.e., *terrible*, *poor*, and *fair*) and favorable (i.e., *good* and *excellent*). The resulting endpoint is then the favorable status. Table 9.2.3 reconstructs a 2×2 table from Table 9.1.6 based on the favorable status. From Table 9.2.3, at baseline, it can be seen that subjects in the placebo group and the test drug group showed similar favorable status with proportions of 45.61% and 42.59%, respectively. However, at visit 3 after treatment, the proportion of subjects with favorable status in the placebo group remained the same while the proportion increased to 72.22% for the test drug group. As a result, it is important to verify whether the proportion of the subjects with favorable status within each group changes significantly after treatment.

For detection of a significant change from the baseline, the same binary endpoint is usually evaluated at the baseline and at various visits after treatment (also see Section 13.5 for McNemar's test). Hence, the resulting binary responses are correlated. In general, data of this type can be summarized by a 2×2 transition table from the baseline as presented in Table 9.2.4. Basically, within each treatment, Table 9.2.4 classifies subjects into four groups: the subjects with favorable status at both the baseline and visit 3 (Y_{11}), the subjects with unfavorable status at both the baseline and visit 3 (Y_{00}), the subjects with a change from unfavorable status at baseline to favorable status at visit 3 (Y_{01}), and the subjects with

Table 9.2.3 Summary of Frequency of Transition of Status from the Baseline at Visit 3

Baseline	Unfavorable	Favorable	Total
<i>Treatment: Placebo</i>			
Unfavorable	23 (40.35%)	8 (14.04%)	31 (54.39%)
Favorable	8 (14.04%)	18 (31.58%)	26 (45.61%)
Total	31 (54.39%)	26 (45.61%)	57 (100.00%)
<i>Treatment: Test Drug</i>			
Unfavorable	14 (25.93%)	17 (31.48%)	31 (57.41%)
Favorable	1 (1.85%)	22 (40.74%)	23 (42.59%)
Total	15 (27.78%)	39 (72.22%)	54 (100.00%)

Note: Unfavorable = terrible, poor, or fair; favorable = good or excellent. See Table 9.1.6.

Table 9.2.4 Summary of Data for Binary Endpoint from the Baseline

Baseline	No	Yes	Total
No	$Y_{00}(P_{00})$	$Y_{01}(P_{01})$	$Y_{0.}(P_{0.})$
Yes	$Y_{10}(P_{10})$	$Y_{11}(P_{11})$	$Y_{1.}(P_{1.})$
Total	$Y_{.0}(P_{.0})$	$Y_{.1}(P_{.1})$	$Y_{..}(1)$

a change from favorable status to unfavorable status at visit 3 (Y_{10}). Hence, Y_{11} and Y_{00} represent the number of the subjects with no change in favorable status, while Y_{10} and Y_{01} are the number of subjects whose favorable status changes from the baseline. Suppose that our objective is to examine whether the proportion of the subjects with favorable status at visit 3 is the same as that at the baseline. In this case, we can compare the number of subjects with favorable status at visit 3 with that at baseline, namely $Y_{.1}$ versus $Y_{1.}$ This type of comparison is known as a *marginal comparison* between visit 3 and the baseline. However, the number of subjects with favorable status at visit 3 is the sum of the number of subjects with the favorable status at both the baseline and visit 3 (Y_{11}) and those with a change from the unfavorable status at baseline to the favorable status at visit 3 (Y_{01}). Similarly, the number of subjects with a favorable status at the baseline is the sum of the number of subjects with favorable status at both the baseline and visit 3 (Y_{11}) and those with a change from the favorable status at the baseline to the unfavorable status at visit 3 (Y_{10}). Consequently, a marginal comparison of the number of subjects with a favorable status between visit 3 and the baseline involves the difference between the number of subjects with a change from the unfavorable status at the baseline to a favorable status at visit 3 and those with a change from the favorable status at the baseline to an unfavorable status at visit 3, namely $Y_{01} - Y_{10}$. The proportion can be obtained by dividing the number of subjects by the sample size.

Statistical hypotheses for comparison of proportions with favorable status between visit 3 and baseline can be expressed as

$$\begin{aligned} H_0: P_{.1} &= P_{1.}, \\ \text{vs. } H_a: P_{.1} &\neq P_{1.}, \end{aligned} \tag{9.2.12}$$

or

$$\begin{aligned} H_0: P_{01} &= P_{10}, \\ \text{vs. } H_a: P_{01} &\neq P_{10}. \end{aligned}$$

Let y_{ij} and p_{ij} denote the corresponding observed number of subjects and proportions for Y_{ij} and P_{ij} , respectively. An unbiased estimator for $\theta = P_{01} - P_{10}$ is given by $p_{01} - p_{10}$. For large samples an estimate of the variance of θ can be obtained as

$$\begin{aligned} v(\theta) &= \text{var}(\theta) \\ &= \frac{1}{n} [(p_{01} + p_{10}) - (p_{01} - p_{10})^2]. \end{aligned} \tag{9.2.13}$$

It follows that a large sample $(1 - \alpha)100\%$ confidence interval for θ is given by

$$(p_{01} - p_{10}) \pm Z(\alpha/2) \sqrt{v(\theta)}. \tag{9.2.14}$$

Under the null hypothesis of $P_{01} = P_{10} = 0$, the estimated large sample variance in (9.2.13) becomes $v(0) = (p_{01} + p_{10})/n$. As a result an asymptotic test statistic for hypotheses (9.2.12) is

$$\begin{aligned}\chi_M^2 &= \frac{(p_{01} - p_{10})^2}{(p_{01} + p_{10})/n} \\ &= \frac{(y_{01} - y_{10})^2}{(y_{01} + y_{10})}.\end{aligned}\tag{9.2.15}$$

We reject the null hypothesis at the α level of significance if

$$\chi_M^2 > \chi^2(\alpha, 1),$$

where $\chi^2(\alpha, 1)$ is the upper α th quantile of a χ^2 random variable with one degree of freedom. However, the test statistics in (9.2.15), which is known as McNemar test (Conover, 1980), is a large-sample approximation of a continuous variable to a discrete variable. χ_M^2 may commit type I error too often with respect to the preselected nominal level when $p_{01} + p_{10} < 0.3$. An approach to overcome this issue is to employ a continuity correction as follows:

$$\chi_M^2 = \frac{(|y_{01} - y_{10}| - 1)^2}{(y_{01} + y_{10})}.\tag{9.2.16}$$

Sometimes, the test statistic with such a continuity correction is too conservative to have sufficient power for detection of a meaningful difference in marginal proportions between visit and baseline. Then, one computes the exact p -value under the null hypothesis. As mentioned before, the comparison in marginal distributions between visit and baseline depends on the subjects with different status at visit and baseline, namely y_{01} and y_{10} . Given the total number of subject with discordant pairs, $y = y_{01} + y_{10}$, both probabilities of a subject with favorable status at the baseline and unfavorable status at the visit and of a subject with unfavorable status at the baseline and favorable status at the visit are equal to 1/2 under the null hypothesis that $P_{01} = P_{10}$. It turns out that y_{01} follows a binomial distribution with a probability of success 0.5 given y . The exact p -value for testing hypotheses (9.2.12) can then be calculated as

$$p\text{-value} = \begin{cases} 2 \sum_{i=1}^{y_{01}} \left[\frac{y!}{i!(y-i)!} \right] (0.5)^y & \text{if } y_{01} < \frac{y}{2}, \\ 2 \left\{ 1 - \sum_{y_{01}=1}^y \left[\frac{y!}{i!(y-i)!} \right] (0.5)^y \right\} & \text{if } y_{01} > \frac{y}{2}, \\ 1 & \text{if } y_{01} = \frac{y}{2}. \end{cases}\tag{9.2.17}$$

The p -values by formula (9.2.17) are calculated by considering the total number of subjects with discordant results to be a fixed known number. Hence this is referred to as the *conditional* exact p -value. Note that the confidence interval and test statistics given in (9.2.14) and (9.2.16) are *unconditional* large-sample procedures. In practice, the number of subjects with different statuses between the visit and the baseline is usually not known until the completion of the trial. It is in fact a random variable and thus technically not a fixed number. As

Table 9.2.5 Results of Analysis in Table 9.2.3

Treatment	Method	Test-Statistic	<i>p</i> -Value
Placebo	Unconditionally large sample		
	No continuity correction	0	1.0000
	With continuity correction	0.0625	0.80259
	Conditional exact		1.0000
Treatment	Unconditionally large sample		
	No continuity correction	14.22	0.00016
	With continuity correction	12.50	0.00041
	Conditional exact		0.00014

an alternative, Suissa and Shuster (1991) propose an unconditional exact procedure for testing hypotheses (9.2.12). The proposed procedure requires a complicated computation of the binomial probability over the range between 0 and 1. However, they provide the critical values for different combinations of significance levels and sample sizes in addition to the required sample sizes for the different significance levels, $P_{01} - P_{10}$, and $P_{01} + P_{10}$.

Example 9.2.2 For the data given in Table 9.2.3, the differences in marginal proportions of favorable status between visit 3 and the baseline for the placebo and the test drug are estimated as 0 and 0.2963, respectively. The 95% large-sample confidence intervals for the difference in proportions for the placebo group and the test drug group are $(-1.038, 1.038)$ and $(0.1641, 0.4284)$, respectively. The results of hypothesis testing by the three procedures discussed above are summarized in Table 9.2.5. All three methods provide consistent results. For the placebo groups, there is no statistically significant difference in proportion of subjects with favorable status between visit 3 and the baseline, while the proportion of the subjects with favorable status at visit 3 for the test drug group is statistically significant higher than that at the baseline at the 5% level of significance.

9.3 INFERENCE OF INDEPENDENT SAMPLES

One of the primary goals of the NINDS trial is to compare the proportion of subjects treated with rt-PA who show early clinical improvement with that of the subjects given a placebo. This type of comparison is usually called a between-group comparison. Comparisons among treatments in randomized, parallel group clinical trials usually involve inference of independent samples because the clinical outcomes of a subject evaluated in terms of efficacy and safety endpoints have nothing to do with those of the subjects within the group or from another group. In this section our efforts will be directed toward statistical inference based on the binary endpoints for two independent samples.

Consider a parallel two group clinical trial that compares a test drug in n_1 subjects with a placebo in n_2 subjects. Table 9.3.1 illustrates a simple way for summarization of a binary endpoint with two categories, *yes* and *no*. *Yes* may mean death, success, improvement, or eradication. In Table 9.3.1, Y_{i1} is the number of subjects with a yes response in the i th group with sample size Y_i , namely $Y_i = n_i$, $i = 1, 2$; and $Y_{..} = n_1 + n_2 = N$ is the total number of subjects in the trial. Since Y_{i1} are independent binomial variables within each group, the population proportion of a *yes* response can be estimated by the methods described in

Table 9.3.1 Data Structure of Binary Endpoint for a Parallel Two-Group Trial

Treatment	Binary Response		Total
	No	Yes	
Test drug	$Y_{10}(P_{10})$	$Y_{11}(P_{11})$	$Y_{1\cdot}(1)$
Placebo	$Y_{20}(P_{20})$	$Y_{21}(P_{21})$	$Y_{2\cdot}(1)$
Total	$Y_{\cdot 0}$	$Y_{\cdot 1}$	$Y_{\cdot \cdot}$

the previous section. However, descriptive measures for comparing proportions of two independent samples often involve differences in proportions, odds ratios, and relative risks.

Let y_{i1} and p_{i1} be the observed number of subjects and proportion with a *yes* response in the i th group. Because of independence, the difference in the population proportions of a *yes* response between two groups can be unbiasedly estimated by

$$\begin{aligned} d_p &= p_{11} - p_{21} \\ &= \frac{y_{11}}{n_1} - \frac{y_{21}}{n_2}, \end{aligned} \quad (9.3.1)$$

with an estimated large-sample variance given as

$$v(d_p) = \frac{p_{11}(1 - p_{11})}{n_1} + \frac{p_{21}(1 - p_{21})}{n_2}. \quad (9.3.2)$$

As a result the lower and upper limits of the large-sample $(1 - \alpha)100\%$ confidence interval for the difference in population proportions can be obtained as

$$d_p \pm Z(\alpha/2)\sqrt{v(d_p)}, \quad (9.3.3)$$

where $Z(\alpha/2)$ is the upper $(\alpha/2)$ th quantile of the normal distribution.

Since treatments can be considered as risk factors, a measure often used to study the relationship between treatments and responses is the *odds ratio*. For example, the NINDS trial defined the odds ratio as the ratio of the odds of a subject with an early clinical improvement after exposure to treatment of rt-PA to the odds of an improvement of a subject given the placebo. The odds of a *yes* response in each treatment group is estimated as

$$\frac{p_{i1}}{p_{i0}} = \frac{y_{i1}}{y_{i0}}.$$

The odds ratio (*OR*) of a *yes* response for comparing the test drug against placebo is then estimated as

$$\begin{aligned} OR &= \frac{p_{11}/p_{10}}{p_{21}/p_{20}} \\ &= \frac{y_{11}/y_{10}}{y_{21}/y_{20}} \\ &= \frac{y_{11}y_{20}}{y_{10}y_{21}}. \end{aligned} \quad (9.3.4)$$

The lower and upper limits of the large-sample $(1-\alpha)100\%$ confidence interval for the odds ratio are given by

$$\exp \left\{ \ln(OR) \pm Z(\alpha/2) \sqrt{\frac{1}{y_{10}} + \frac{1}{y_{11}} + \frac{1}{y_{20}} + \frac{1}{y_{21}}} \right\}, \quad (9.3.5)$$

where $Z(\alpha/2)$ is the upper $(\alpha/2)$ th quantile of the normal distribution; “exp” and “ln” denote the natural exponential and logarithm, respectively.

Another commonly employed measure for describing the relationship between treatment and its associated binary outcomes is the relative risk. The risk for obtaining a yes response for the i th group is simply the proportion of the subjects with a yes response. As a result, the relative risk (RR) for the test drug to placebo can be estimated by

$$RR = \frac{p_{11}}{p_{21}}. \quad (9.3.6)$$

An estimated large-sample variance for the logarithm of the relative risk is given as

$$v[\ln(RR)] = \frac{1-p_{11}}{y_{11}} + \frac{1-p_{21}}{y_{21}}. \quad (9.3.7)$$

It follows that the large-sample $(1-\alpha)100\%$ confidence interval for the relative risk are given as

$$\exp \left\{ \ln(RR) \pm Z(\alpha/2) \sqrt{\frac{1-p_{11}}{y_{11}} + \frac{1-p_{21}}{y_{21}}} \right\}. \quad (9.3.8)$$

Note that the difference in proportion is an adequate measure for comparing two proportions when they are not close to 0, and when Y_{ij} is greater than 5, for $j = 0, 1$, and $i = 1, 2$. This measure that is symmetric about 0. On the other hand, when the proportion of a yes response is close to 0, then either the odds ratio or the relative risk is a better measure for comparing two proportions than the difference. Since the odds ratio and the relative risk range from 0 to infinity, the distributions of their estimates are skewed. If there is no difference in true proportions of a yes response between two population groups, the odds ratio or relative risk can be expected to be close to 1. When the number of subjects in each group becomes large, the estimated odds ratio and relative risk will be very close to each other. In clinical trials the difference in proportions, odds ratio, and relative risk are useful measures for comparing prospectively two independent proportions. Note that the odds ratio can also be employed for retrospective case control studies in which the difference in proportions and relative risk is inappropriate.

Example 9.3.1 For the NINDS trial, estimates of the subjects experiencing early improvement for each group are given in Table 9.2.2. From these sample proportions, the difference in population proportions of the subjects with early improvement between rt-PA and the placebo is estimated by 8.01% with an estimated large-sample variance of 0.0016. This leads to a 95% confidence interval of (0.17%, 15.76%). The results are summarized in Table 9.3.2. As a result, on average, the proportion of the subjects treated with rt-PA showing an early improvement is about 8% higher than those given a placebo. Note that the 95% confidence interval does not contain zero, which indicates that significantly more subjects receiving rt-PA can show early improvement than those receiving the placebo.

For the U.S. Physicians' Health Study, the proportion of subjects with development of malignant neoplasm after an average of 12 years treatment is 11.53% for subjects receiving

Table 9.3.2 Comparison of Subjects with Improvement in the NINDS Trial

Treatment	<i>N</i>	Improvement	Difference (<i>se</i>)	95% Confidence Interval
rt-PA	312	147 (47.12%)	0.0801 (0.04)	(0.0017, 0.1576)
Placebo	312	122 (39.10%)		

bet a carotene and 11.72% for those receiving a placebo. The unadjusted odds ratio of malignant neoplasm for the subjects treated with rt-PA as compared to the placebo is 0.98 with a 95% confidence interval of (0.91, 1.07) (see Table 9.3.3). Similarly, the estimate of unadjusted relative risk is 0.98, and its corresponding large-sample confidence interval is from 0.92 to 1.06. The estimated odds ratio and relative risk and their associated 95% confidence interval are consistent and numerically close. Since both 95% confidence intervals for the odds ratio and relative risk include 1, we conclude that the long-term treatment of beta carotene does not statistically significantly reduce odds or risk of development of malignant neoplasm at the 5% level of significance.

Statistical inference of binary data from a parallel two group trial involves testing the hypotheses regarding the difference in proportions with a *yes* response between the test drug group and the placebo group. This can be expressed as

$$\begin{aligned} H_0: P_{11} &= P_{21}, \\ \text{vs. } H_a: P_{11} &\neq P_{21}. \end{aligned} \quad (9.3.9)$$

For testing hypotheses (9.3.9), several statistical procedures are available in the literature. Among these, the method using the difference in proportions is probably the most commonly used. Under the null hypothesis the proportion of the subjects with a *yes* response is assumed to be equal. As a result we can combine both groups to estimate of the common proportion of the subjects with a *yes* response. This gives

$$p_0 = \frac{y_{11} + y_{21}}{n_1 + n_2} \quad (9.3.10)$$

with an estimated large-sample variance

$$v(p_0) = \frac{p_0(1 - p_0)(n_1 + n_2)}{n_1 n_2}. \quad (9.3.11)$$

A test statistic for hypothesis (9.3.9) is the usual Z-statistic:

$$Z = \frac{P_{11} - P_{21}}{\sqrt{v(p_0)}}. \quad (9.3.12)$$

Table 9.3.3 Summary of the Estimated Odds Ratio and the Relative Risk for Malignant Neoplasm due to Beta Carotene in U.S. Physicians' Health Study

Beta Carotene (<i>N</i> =11,036)	Placebo (<i>N</i> =11,036)	Odds Ratio (95% CI)	Relative Risk (95% CI)
1273 (11.53%)	1293 (11.72%)	0.98 (0.91, 1.07)	0.98 (0.92, 1.06)

We then reject the null hypothesis and conclude that the proportion of the subjects receiving the test drug with a *yes* response is statistically significantly different from that for subjects receiving the placebo at the α level of significance if $|Z| > z(\alpha/2)$, where $Z(\alpha/2)$ is the upper $(\alpha/2)$ th quantile of the standard normal distribution.

Another commonly employed method is the chi-square test which is a useful statistical method for examining the relationship between two factors. In the context of comparing the proportions between two treatments, one factor is the treatment with two levels (i.e., the active test drug and the placebo), and the other factor is the binary clinical endpoint with two categories (i.e., a *yes* response and a *no* response). For the 2×2 contingency table given in Table 9.3.1, the chi-square test for testing independence between the two factors is equivalent to testing the difference in proportion of the subjects with a *yes* response between two groups. Under the assumption of independence, the marginal total (i.e., the number of subjects in each group $Y_{1.}$ and $Y_{2.}$) and the total number of the subjects in each category of the binary clinical endpoint (i.e., $Y_{.0}$ and $Y_{.1}$) are considered fixed. The expected frequency for the (i, j) cell is then given by

$$m_{ij} = \frac{(y_{i.})(y_{.j})}{N}, \quad i = 1, 2, \quad j = 0, 1. \quad (9.3.13)$$

If the response has nothing to do with the treatment, we can expect the observed frequencies to be fairly close the expected frequencies. As a result the weighted sum of squares of the differences between the observed and expected frequencies, with expected frequencies as weights, can serve as a test statistic for comparing two independent proportions

$$\chi_P^2 = \sum_{i=1}^2 \sum_{j=0}^1 \frac{(y_{ij} - m_{ij})^2}{m_{ij}}. \quad (9.3.14)$$

The above statistic χ_P^2 is usually referred to as Pearson's chi-square test, and its computation is given in Table 9.3.4. Under the null hypothesis of no difference in proportions, χ_P^2 follows a χ^2 distribution with one degree of freedom if the sample size of each group is moderate (i.e., greater than 30) and the expected frequency in each cell is at least 5. We reject the null hypothesis in (9.3.9) if $\chi_P^2 > \chi^2(\alpha, 1)$, where $\chi^2(\alpha, 1)$ is the upper α th

Table 9.3.4 Computation of χ_P^2 Statistics for Hypothesis (8.3.6) Based on Binary Data for a Parallel Two-Group Trial

Treatment	Binary Response		Total
	No	Yes	
Test drug			
<i>O</i>	y_{10}	y_{11}	$y_{1.}$
<i>E</i>	m_{10}	m_{11}	
$O - E$	$y_{10} - m_{10}$	$y_{11} - m_{11}$	
$(O - E)^2/E$	$(y_{10} - m_{10})^2/m_{10}$	$(y_{11} - m_{11})^2/m_{11}$	
Placebo			
<i>O</i>	y_{20}	y_{21}	$y_{2.}$
<i>E</i>	m_{20}	m_{21}	
$O - E$	$y_{20} - m_{20}$	$y_{21} - m_{21}$	
$(O - E)^2/E$	$(y_{20} - m_{20})^2/m_{20}$	$(y_{21} - m_{21})^2/m_{21}$	
Total	$y_{.0}$	$y_{.1}$	$y_{..}$

quantile of a chi-square random variable with one degree of freedom. Note that the test can be modified with a continuity correction as follows:

$$\chi_{PC}^2 = \sum_{i=1}^2 \sum_{j=0}^1 \frac{\{\max[0, |y_{ij} - m_{ij}| - 0.5]\}^2}{m_{ij}}. \quad (9.3.15)$$

The above chi-square test is easy to employ. In addition, it can be extended to the situation where two factors have r and c categories, respectively. The expected frequency for the (i, j) th cell can be calculated by the formula provided in (9.3.13) as the product of marginal frequencies for the i th level of one factor and the j th level of the other divided by the total sample size. The test statistics χ_P^2 now is the weighted sum of squares over all $r \times c$ cells. The test for independence can be similarly performed by comparing χ_P^2 with the upper α th quantile of a chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

When sample size is small or the expected frequency in each cell is smaller than 5, approximation by the large-sample procedures such as Z-statistic or chi-square test may be inadequate. In this case, we can consider the exact method for inference of comparing two proportions. Under the assumption that the marginal frequencies are fixed, there is only one cell frequency in the 2×2 contingency table given in Table 9.3.1, which is allowed to vary. If we allow the number of subjects treated with test drug having a *yes* response, Y_{11} , to vary, then Y_{11} follows a hypergeometric distribution with the probability for observing frequency y_{11} given as follows:

$$P\{Y_{11} = y_{11}\} = \frac{(y_{11})!(y_{20})!(y_{10})!}{N!(y_{10})!(y_{20})!(y_{11})!(y_{21})!}. \quad (9.3.16)$$

It follows that the two-tailed p -value for the Fisher's exact test is the sum of probabilities over a set of tables with $P(Y_{11})$ less than or equal to the probability calculated from the observed frequency y_{11} .

Note that the only requirement for the assumption of a hypergeometric distribution in (9.3.16) and the expected frequency of Y_{11} in (9.3.13) is randomization of subjects to receive either the test drug or the placebo. As a result the variance of Y_{11} derived from hypergeometric distribution is given as

$$v_{11} = \frac{y_{11}y_{20}y_{10}y_{11}}{N^2(N-1)}. \quad (9.3.17)$$

The randomization chi-square test statistic is then given by

$$\chi_R^2 = \frac{(y_{11} - m_{11})^2}{v_{11}}. \quad (9.3.18)$$

Koch and Edwards (1988) show that the relationship between the Pearson's and randomization chi-square test statistics can be expressed as

$$\chi_R^2 = \frac{N-1}{N} \chi_P^2. \quad (9.3.19)$$

Example 9.3.2 We continue to use the binary endpoint of early clinical improvement from the NINDS trial for illustration of the statistical testing procedures discussed above. Under

the null hypothesis on the equal proportion of subjects with early clinical improvement in both groups, the common proportion is estimated as $(122 + 147)/[2(312)] = 43.11\%$ with an estimated large-sample variance of 0.001572. The Z-statistic is then given by $(0.4712 - 0.3910)/\sqrt{0.001572} = 2.0209$ which is greater than $Z(0.025) = 1.96$. The corresponding two-tailed p -value is 0.043, and we conclude that the proportion of the subjects treated with rt-PA showing an early clinical improvement is statistically significantly greater than that given placebo. Since the sample size of both groups is the same (312), the expected frequency of an early improvement is the same for both groups, which equals $(269)(312)/624 = 134.5$. Similarly, the expected frequency of no improvement is 177.5 for both groups. The sum of all four expected frequencies is equal to the total sample size of 624. The test statistic χ^2_p without the continuity correction, computed according to (9.3.14), is 4.084 which is greater than $\chi^2(0.05, 1) = 3.84$. The corresponding p -value is 0.043. Also, randomization of the chi-square gives

$$\chi^2_R = (623/624)4.084 = 4.077$$

with a p -value of 0.043. However, the observed value of χ^2_{PC} with continuity correction from (9.3.15) is 3.764 which is less than 3.84. The corresponding two-tailed p -value is 0.052. The Fisher's exact test also gives a two-tailed p -value of 0.052. Consequently, according to the chi-square test with the continuity correction or the Fisher's exact test, we fail to reject the null hypothesis of equal proportions of an early improvement for both groups. Therefore this numerical example provides an example in which the choice of continuity correction or exact test can reach different conclusion from those by large-sample approximation methods in rejection of the null hypothesis.

9.4 ORDERED CATEGORICAL DATA

In clinical trials it is not uncommon to have more discrete efficacy and safety endpoints with more than two categories. For example, although an adverse event can be classified into dichotomous groups, such as serious or nonserious adverse event, as a binary response, the intensity of an adverse event is evaluated according to the categories *mild*, *moderate*, or *severe*. Similar examples are seen in laboratory safety assessments. For example, we can assess the safety of a drug based on a particular laboratory parameter such as aspartate transaminase (AST) or alanine transaminase (ALT). Using the observed value of this particular parameter, subjects are usually classified into three categories: *below*, *within*, and *above* the referenced laboratory normal range. For another example, each individual symptom score in the composite AUA symptom index for benign prostatic hyperplasia (BPH) presented in Table 9.1.3 consists of six categories. These six categories evaluate individual symptoms in an increasing severity, from *not at all* as the first category to *almost always* as the last category. Although severity of a symptom is actually a continuous variable, it is, however, extremely difficult or impractical to measure the severity as a continuous variable objectively. Therefore it is an unobserved latent variable. In practice, the continuous spectrum of severity for a symptom is divided into several ordinal categories to objectively evaluate the symptom. This type of categorical data is called the *ordered categorical data*. Tables 9.1.5 and 9.1.6 provide other examples of ordered categorical data in the areas of diagnostic imaging and respiratory disorder.

Table 9.4.1 Data Structure of Polychotomous Categorical Endpoints for a Parallel Two-Group Trial

Treatment	Category				Total
	1	2	...	J	
Test drug	$Y_{11}(P_{11})$	$Y_{12}(P_{12})$...	$Y_{1J}(P_{1J})$	$Y_{1\cdot}(1)$
Placebo	$Y_{21}(P_{21})$	$Y_{22}(P_{22})$...	$Y_{2J}(P_{2J})$	$Y_{2\cdot}(1)$
Total	$Y_{\cdot 1}$	$Y_{\cdot 2}$...	$Y_{\cdot J}$	N

The data structure of polychotomous categorical data from a parallel two-group trial is provided in Table 9.4.1, where Y_{ij} represents the number of subjects in the j th category for the i th group, $n_i = Y_{i\cdot}$, $j = 1, \dots, J$, $i = 1, 2$. It is then of interest to evaluate whether the distribution of the subjects across categories is the same for the test drug and the placebo. For example, if the test drug is effective in treatment of subjects with benign prostatic hyperplasia, then we would expect a shift in distribution such that more subjects are in the categories *not at all*, *less than 1 times in 5*, or *less than half the time*. As a consequence of randomization (nonadaptiveness) of subjects to treatments, the null hypothesis of interest can be formulated as follows:

$$H_0: \text{The distribution of subjects in response categories is the same as for both groups.} \quad (9.4.1)$$

This hypothesis in fact tests whether there is relationship between the treatment and response categories. Under the null hypothesis and a randomization structure, Y_{ij} follow a hypergeometric distribution

$$P\{Y_{ij}\} = \frac{\prod_{i=1}^2 (Y_i)! \prod_{j=1}^J (Y_j)!}{N! \prod_{i=1}^2 \prod_{j=1}^J (Y_{ij})!}, \quad j = 1, \dots, J, i = 1, 2. \quad (9.4.2)$$

Similarly to the 2×2 table in (9.3.13), the expected frequencies of Y_{ij} is the product of marginal frequencies in the j th category and the sample size of the i th group divided by the total sample size, namely

$$m_{ij} = \frac{y_{i\cdot} y_{\cdot j}}{N}, \quad i = 1, 2, j = 1, \dots, J. \quad (9.4.3)$$

Substitution of y_{ij} and m_{ij} into (9.3.14) and summation over all $2J$ cells gives the Pearson's chi-square test statistic, which has a χ^2 distribution with $J - 1$ degrees of freedom if sample size of each group is moderate (> 30) and the expected frequency in each cell is at least 5. To obtain the randomization chi-square statistic, we need to know the covariance between Y_{ij} and $Y_{i'j'}$, which is given by

$$\text{cov}(Y_{ij}, Y_{i'j'}) = \frac{m_{ij}(Nd_{ii'} - y_{i\cdot})(Nd_{jj'} - y_{\cdot j})}{N(N-1)}, \quad (9.4.4)$$

where $d_{ii'} = 1$, if $i = i'$ and $d_{ii'} = 0$, if $i \neq i'$; and $d_{jj'} = 1$, if $j = j'$; and $d_{jj'} = 0$, if $j \neq j'$.

Let \mathbf{y} , \mathbf{m} , and \mathbf{V} denote the vectors of the observed and expected frequencies and the covariance matrix of \mathbf{y} , respectively. Also, let $\mathbf{A} = [\mathbf{I}_{J-1}, \mathbf{0}_{(J-1) \times 2J}]$, where \mathbf{I}_{J-1} is a $(J-1) \times (J-1)$ identity matrix and $\mathbf{0}_{(J-1) \times 2J}$ is a $(J-1) \times 2J$ matrix of 0's. Koch et al. (1982) show that the randomization chi-square can be expressed as

$$\begin{aligned}\chi^2_R &= (\mathbf{y} - \mathbf{m})' \mathbf{A}' (\mathbf{A} \mathbf{V} \mathbf{A}')^{-1} \mathbf{A} (\mathbf{y} - \mathbf{m}) \\ &= \frac{N-1}{N} \chi^2_P.\end{aligned}\quad (9.4.5)$$

Both the Pearson's and randomization chi-square tests are useful for detection of the existence of general association between treatment and categorical response in either the nominal or ordinal scale. However, they cannot identify a particular relationship. Suppose that the categorical data are ordinal, one might want to see whether there is a difference in the central tendency of the distributions between two treatments. In other words, we may want to detect a location shift in response categories. Let $\mathbf{a} = (a_1, a_2, \dots, a_J)$ be a set of scores to reflect the ordinal nature of response categories. Then the mean score for the i th group can be computed in the usual manner as we calculate the sample mean for a frequency table:

$$c_i = \sum_{j=1}^J \frac{a_j y_{ij}}{y_{i\cdot}}, \quad i = 1, 2. \quad (9.4.6)$$

Because there are only two treatment groups and all marginal frequencies are assumed to be fixed, the mean score for only one treatment group is random. Hence we can only consider the mean score for the test drug. Koch and Edwards (1988) give the expected value and variance of c_1 , respectively, as

$$c = \sum_{j=1}^J \frac{a_j y_{j\cdot}}{N}$$

and

$$v(c_1) = \frac{(N - y_{i\cdot}) v_a}{(N-1)y_{1\cdot}}, \quad (9.4.7)$$

where

$$v_a = \sum_{j=1}^J \frac{(a_j - c)^2 y_{j\cdot}}{N}, \quad i = 1, 2.$$

From (9.4.6) and (9.4.7), the Z-statistic can be constructed to test whether there is a location shift in distribution between the two groups:

$$Z = \frac{c_1 - c}{\sqrt{v(c_1)}},$$

or equivalently

$$X = Z^2 = \frac{(c_1 - c)^2}{v(c_1)}. \quad (9.4.8)$$

If the sample size is fairly large, X approximately follows a chi-square variable with one degree of freedom. As a result we reject the null hypothesis of no location shift between two distributions at the α level of significance if $X > \chi^2(\alpha, 1)$.

The large-sample randomization test statistic for location shift can be expressed as

$$X = \frac{K(c_1 - c_2)^2}{v_a[(1/y_{1.}) + (1/y_{2.})]}, \quad (9.4.9)$$

where $K = (N - 1)/N$ is the finite population correction factor.

Note that except for the finite population correction factor, X is in fact a test statistic for the detection of mean difference between the two independent samples. As a result, as in the relationship between t statistic and F statistic for continuous endpoints, the test statistics X for two independent samples given in (9.4.9) can be easily extended to test overall location shift for I independent samples, $I > 2$. The resulting test statistic is the one-way analysis of variance based on scores $\mathbf{a} = (a_1, a_2, \dots, a_I)$ which is given as

$$X = (N - 1) \sum_{i=1}^I \frac{y_i(c_i - c)^2}{v_a}, \quad (9.4.10)$$

where c_i and v_a are similarly defined as in (9.4.6). Note that except for the constant $(N - 1)$, the numerator in (9.4.10) is the between-group sum of squares, and Nv_a represents the total sum of squares. If the sample size in each treatment group is at least moderate, X approximately follows a chi-square variable with $I - 1$ degrees of freedom.

The choice of ordinal scores affects the statistical analysis and its clinical interpretation, which usually depends on the question being asked. For example, the last question in the AUA-7 symptom index for BPH is on *how many times a subject gets up during the night to urinate*. The symptom score for this question is a discrete count such as integers 0, 1, 2, ..., and J . The scores of this kind are called integer scores. If categorical clinical endpoints represent some discrete counts or the classes are equally spaced, one might consider the use of integer scores. Alternative scores suggested by Koch and Edwards (1988) are standardized midrank scores and logrank scores. Standardized midrank scores are the expected values of order statistics of uniform distribution. Koch and Bhapkar (1982) showed that the resulting test statistic is equivalent to the Wilcoxon rank sum test (or Kruskal-Wallis test for $I > 2$) without scaling the categories. Standardized midrank scores are also referred to as modified ridit scores (van Elteren, 1960; Lehmann, 1975). They can be assessed by means of PROC FREQ of SAS® with the statement MIDRIDIT in the option SCORES. Alternatively, one can use logrank scores or Savage scores, which are the expected values of order statistics from an exponential distribution. These scores are useful when the interest is to detect treatment difference for a distribution of the data that is L-shaped. The logrank scores can be generated by option statement of SAVAGE in PROC NPAR1WAY of SAS®.

Note that test statistic X for detecting a location shift has fewer degrees of freedom than the randomization test, χ^2_R or Pearson's chi-squares test, χ^2_P for the general association. In general, X is a more powerful test than either χ^2_R or χ^2_P . The approximation of χ^2_R or χ^2_P by chi-square distribution depends on cell frequencies y_{ij} . Basically the test statistics for the detection of a location shift can be considered a linear rank statistics that is a linear combination of scores. Hence, their large-sample approximation depends on a linear combination of y_{ij} . Therefore, the sample size requirement for X for detecting a location shift is less stringent than that of χ^2_R or χ^2_P .

Table 9.4.2 Summary of Analysis of Validation for Question 4–Frequency No. 1 in AUA Symptom Index

Category	BPH	Control	χ_P^2	χ_R^2	X
N	73	56			
Not at all	8 (10.96%)	16 (28.57%)	16.87	16.74	14.9
Less than 1 in 5 times	18 (24.66%)	23 (41.07%)	(0.0047)	(0.0050)	(0.0001)
Less than half the time	17 (23.29%)	7 (12.50%)			
About half the time	14 (19.18%)	7 (12.50%)			
More than half the time	8 (10.96%)	2 (3.57%)			
Almost always	8 (10.96%)	1 (1.79%)			

Example 9.4.1 As indicated by Barry et al. (1992), the AUA symptom index should be able to discriminate the subjects from BPH from those without BPH. The data given in Table 9.1.4 summarize the results of the initial question 4 for urination frequency in two hour since the last urination in a validation study from the subjects with BPH and controls for the AUA symptom index. Pearson's chi-square is 16.874 with a p-value of 0.0047 obtained from a chi-square random variable with five degrees of freedom. The randomization chi-square statistic for general association can then be obtained by the relationship

$$\begin{aligned}\chi_R^2 &= \frac{N - 1}{N} \chi_P^2 \\ &= (128/129)16.874 \\ &= 16.743\end{aligned}$$

with a corresponding p -value of 0.005. Both Pearson's and the randomization chi-square tests reject the null hypothesis at the 5% level of significance, that the response distribution with respect to urination frequency is the same as for the subjects with BPH and normal controls. However, this general association does not answer the question of whether urination frequency for the subjects with BPH is higher than normal controls. From Table 9.4.2, 41.1% of the subjects with BPH had to urinate in less than two hours since the last urination about half the time, more than half the time, or almost always, while the proportion is only 17.86% for controls. Therefore, there appears to be a location shift in the distributions of categories between the two groups. Since the AUA symptom index suggests the use of integers from 0 to 5 for each of seven symptom scores, we computed the test statistic with integers scores 0, 1, 2, 3, 4, 5 for detection of location shift which turns out to be 14.90 with a p -value of 0.0001. This result confirms our suspicion that the subjects with BPH tend to urinate more frequently in two hours of the last urination than controls.

9.5 COMBINING CATEGORICAL DATA

In multicenter clinical trials, the studies are conducted at different centers at which randomization schedules are independently generated. In addition, subjects may be stratified based on some demographic and/or patient baseline characteristics such as gender, race, or

Table 9.5.1 Data Structure of Binary Endpoint with H Strata for a Parallel Two-Group Trial

Treatment	Binary Response		Total
	No	Yes	
Test drug	$Y_{h10}(P_{h10})$	$Y_{h11}(P_{h11})$	$Y_{h1..}(1)$
Placebo	$Y_{h20}(P_{h20})$	$Y_{h21}(P_{h21})$	$Y_{h2..}(1)$
Total	$Y_{h..0}$	$Y_{h..1}$	$Y_{h..} = N_h$

Note: $h = 1, \dots, H$.

the severity of disease. The NINDS trial is a typical clinical trial of this kind. In addition the NINDS trial is divided into two parts with additional stratification with respect to the time from the onset of the stroke to the start of treatment. In practice, the treatment imbalance would occur by chance with respect to some important covariates that are not used as stratified factors in the randomization process. These covariates are usually called the post-randomization stratified factors. As a result, we can compare the differences between the test drug and the placebo by combining the results from a set of strata with an appropriate adjustment of the effects caused by the covariates.

Table 9.5.1 illustrates binary response data obtained from an H strata for comparing a test drug with a concurrent placebo group. If all marginal frequencies are considered fixed, similar to (9.3.13) and (9.3.17), the expected frequency and variance for the h th strata are given by, respectively,

$$m_{h11} = \frac{y_{h1..} y_{..1}}{N_h},$$

and

$$v_h = \frac{y_{h1..} y_{h2..} y_{..0} y_{..1}}{N_h^2(N_h - 1)}, \quad h = 1, \dots, H. \quad (9.5.1)$$

Mantel and Haenszel (1959) propose the well-known Mantel-Haenszel (MH) statistic to combine the results of the difference in proportions of a *yes* response from H different strata. The calculation of the MH statistic is illustrated in Table 9.5.2. After the expected frequency and variance of Y_{11} are calculated for each stratum, the sum of the differences between the observed and expected frequencies and the sum of variances over all strata can

Table 9.5.2 Summary of Results of Binary Endpoint with H Strata

Strata	Observed Frequency	Expected Frequency	Difference	Variance
1	y_{111}	$m_{111} = y_{11..} y_{1..1} / N_1$	$y_{111} - m_{111}$	v_1
:	:	:	:	
h	y_{h11}	$m_{h11} = y_{h1..} y_{h..1} / N_h$	$y_{h11} - m_{h11}$	v_h
:	:	:	:	
H	y_{H11}	$m_{H11} = y_{H1..} y_{H..1} / N_H$	$y_{H11} - m_{H11}$	v_H
Sum	$\sum y_{h11}$	$\sum m_{h11}$	$\sum (y_{h11} - m_{h11})$	$\sum v_h$

Note: $v_h = [y_{h1..} y_{h2..} y_{..0} y_{..1}] / [N_h^2(N_h - 1)]$, $h = 1, \dots, H$.

be obtained. The MH statistic is the ratio of the square of the sum of the differences to the sum of variances:

$$X_{\text{MH}} = \frac{\left[\sum_{h=1}^H (y_{h11} - m_{h11}) \right]^2}{\sum_{h=1}^H v_h}. \quad (9.5.2)$$

When the sum of sample size over all strata is sufficiently large, then X_{MH} approximately follows a chi-square distribution with one degree of freedom. Consequently, we reject the null hypothesis of no difference between the test drug and the placebo after adjustment of covariates at the α level of significance if

$$X_{\text{MH}} > \chi^2(\alpha, 1).$$

The numerator in the MH statistic is the square of the sum of the differences between the observed and expected frequencies. Therefore, the MH statistic would be more powerful if the association between treatment and response were in the same direction over all the strata. On the other hand, if the difference between the test drug and the placebo is not homogeneous across strata, the differences in (9.5.2) will cancel each other out, and resulting sum will be small. Therefore, the MH statistic is not powerful in the presence of heterogeneous treatment across strata in the same direction. A test statistic for the detection of variation of treatment effects across the strata is

$$X_V = X_T - X_{\text{MH}}, \quad (9.5.3)$$

where

$$X_T = \frac{1}{v_h} \sum_{h=1}^H (y_{h11} - m_{h11})^2. \quad (9.5.4)$$

One can test the null hypothesis for the absence of heterogeneity in the treatment effect by comparing the observed X_V with a chi-square random variable with $H - 1$ degrees of freedom. If there is no evidence for the presence of variation of treatment differences across strata, Yusuf et al. (1985) suggests an estimate for the common odds ratio of a yes response as

$$OR_C = \exp \left\{ \frac{\sum_{h=1}^H (y_{h11} - m_{h11})}{\sum_{h=1}^H v_h} \right\}, \quad (9.5.5)$$

with the corresponding lower and upper limits of a $(1 - \alpha)100\%$ confidence interval for the population common odds ratio as follows:

$$\exp \left\{ \frac{\sum_{h=1}^H (y_{h11} - m_{h11})}{\sum_{h=1}^H v_h} \pm \frac{Z(\alpha/2)}{\sqrt{\sum_{h=1}^H v_h}} \right\}. \quad (9.5.6)$$

Example 9.5.1 To illustrate the MH method for combining treatment effects, we use the data from the NINDS trial. Since the trial has two parts and the time was stratified into two intervals: 0 to 90 and 91 to 180 minutes, there are four 2×2 tables, one for each strata. The frequency used for calculation of the MH test is the number of subjects treated with rt-PA who showed clinical improvement. Table 9.5.3 gives the observed and expected frequencies, the difference, variance, and the square of the difference divided by variance.

Table 9.5.3 Comparison Between rt-PA and a Placebo Effect in Subjects Showing Clinical Improvement for Adjustment of Part and Time from Onset to Treatment in the NINDS Trial

Part	Time	Observed Frequency	Expected Frequency	$O_h - E_h$	v_h	$(O_h - E_h)^2 / v_h$
1	0–90	36	34.2230	1.7770	8.7351	0.3615
1	91–180	31	27.3750	3.6250	8.9513	1.4680
2	0–90	51	42.7362	8.2638	10.2188	6.6829
2	91–180	29	30.8706	-1.8706	10.0230	0.3491
Sum		147	135.2048	11.7952	37.9281	8.8615

Note: Observed and expected frequencies are for those subjects showing an improvement in rt-PA group.

The sum of the differences between the observed and expected frequencies is 11.7952 and the sum of variances is 37.9281. Thus, by (9.5.2), the MH statistic is given by

$$X_{\text{MH}} = \frac{(11.7952)^2}{37.9281} = 3.6682$$

with a *p*-value of 0.055. Therefore, we fail to reject the null hypothesis of no treatment effect at the 5% significance level. From Table 9.5.3, X_T is given by 8.8615. As a result,

$$X_V = 8.8615 - 3.6682 = 5.1933$$

which is smaller than $\chi^2(0.05, 3) = 7.81$. Therefore, there is no strong evidence to doubt the presence of significant variation of treatment effects across strata. However, from the fifth column in Table 9.5.3, there seems to be some heterogeneity of differences among strata. The first three strata indicates that odds of clinical improvement for subject treated with rt-PA is greater than those given placebo, while the strata for the combination of part 2 and time from 91 to 180 minutes gave the result in the opposite direction. However, only the stratum with the combination of part 2 and time from 0 to 90 minutes provided a statistical significant treatment effect at the 5% level of significance. The common odds ratio is then estimated by

$$\exp\left(\frac{11.7952}{37.9281}\right) = \exp(0.3110) = 1.3648$$

with the corresponding 95% confidence interval

$$\exp\left(0.3110 \pm \frac{1.96}{\sqrt{37.9281}}\right) = (0.9928, 1.8761).$$

Note that the 95% confidence interval contains 1. This result is consistent with that from hypothesis testing that no statistically significant treatment effect was detected at the 5% level of significance.

Mantel (1963) showed that similar techniques can be extended to combine the results of discrete endpoints with a total of J ordered categories from H strata in clinical trials comparing a test drug with a placebo group. Let Y_{hij} represent the number of subjects in the j th category for the i th group from the h th stratum, $n_i = Y_{i..}$, $j = 1, \dots, J$, $i = 1, 2$, $h = 1, \dots, H$. Since subjects at different strata are different, any clinical endpoints observed from subjects at different strata are statistically independent. Therefore, the probability of

Y_{hij} follows a product hypergeometric distribution given by

$$P\{Y_{hij}\} = \prod_{h=1}^H \left\{ \frac{\prod_{i=1}^2 (Y_{hi.})! \prod_{j=1}^J (Y_{hj})!}{(N_h)! \prod_{i=1}^2 \prod_{j=1}^J (Y_{hij})!} \right\}. \quad (9.5.7)$$

Let c_{h1} , c_h , and $v(c_{h1})$ be the mean score, expected value, and variance for the h th strata as defined in (9.4.6) and (9.4.7), respectively. Then, the extended Mantel-Haenszel statistic for detection of a location shift in distribution between the test drug and the placebo across strata is given by

$$X_{EMH} = \frac{\left[\sum_{h=1}^H (c_{h1} - c_h) \right]^2}{\sum_{h=1}^H v(c_{h1})}. \quad (9.5.8)$$

One can reject the null hypothesis of no location shift after adjustment for covariates at the α level of significance if

$$X_{EMH} > \chi^2(\alpha, 1).$$

If one chooses to use standard midrank scores, then X_{EMH} is the same as the test procedure proposed by van Elteren (1960) for combining Wilcoxon rank sum statistics over a set strata, which is also referred to as the blocked Wilcoxon rank sum statistic (Lehmann, 1975). Extension to more than two treatment groups and a test for homogeneity of location shifts across strata and related issues can be found in Koch and Edwards (1988).

Similar to the Mantel-Haenszel estimator for the average odds ratio for the stratified 2×2 table given in (9.5.5), Davis and Chung (1995) suggest an MH estimator, originally proposed by Mantel (1963), to estimate the common average treatment effect across strata. Let c_{hik} be the resulting observation of the k th subject in the i th treatment form the h th strata after application of scores, a_1, a_2, \dots, a_J to each subject. Also let c_{hi} and s_{hi}^2 be the mean score and sample variance for treatment i and stratum h . Then we have

$$c_{hi} = \frac{1}{y_{hi.}} \sum_{k=1}^{Y_{hi.}} c_{hik},$$

and

$$s_{hi}^2 = \frac{1}{y_{hi.}-1} \sum_{k=1}^{Y_{hi.}} (c_{hik} - c_{hi})^2. \quad (9.5.9)$$

The MH estimator of the average treatment effect proposed by Davis and Chung (1995) is given as

$$\hat{\theta}_{MH} = \frac{\sum_{h=1}^H w_h (c_{h1} - c_{h2})}{\sum_{h=1}^H w_h}, \quad (9.5.10)$$

where

$$w_h = \frac{y_{h1} \cdot y_{h2}}{N_h}, \quad h = 1, \dots, H.$$

Therefore $\hat{\theta}_{MH}$ is a weighted average of the stratum-specific treatment differences with weight w_h . Davis and Chung (1995) also propose the following estimate for the variance of $\hat{\theta}_{MH}$:

$$\hat{v}(\hat{\theta}_{MH}) = \frac{1}{\left[\sum_{h=1}^H w_h \right]^2} \left\{ \sum_{h=1}^H w_h^2 \left[\frac{s_{h1}^2}{y_{h1}} + \frac{s_{h2}^2}{y_{h2}} \right] \right\}. \quad (9.5.11)$$

Table 9.5.4 Estimation of Average Treatment Difference Based on the Change in Diagnostic Confidence Given in Table 9.1.6

	Stratum	
	Female	Male
<i>Mean Score</i>		
Test	5.13889	5.05405
Placebo	3.32432	3.33333
Difference (D)	1.81457	1.72072
Weight (W)	18.2466	18.2466
$W * D$	33.1096	31.3973
<i>Variance</i>		
Test	1.49444	1.60811
Placebo	2.66976	2.51429
$W^2 * \text{sum of variance of } y_{hi}$	37.8475	37.7230

They indicate that $\hat{\theta}_{\text{MH}}$ is equivalent to that derived under the fixed-effects analysis of variance estimator from the main effects model. However, $\hat{\theta}_{\text{MH}}$ and $\hat{v}(\hat{\theta}_{\text{MH}})$ are obtained only under the assumption of randomization of subjects to the treatment assignment without the assumptions of normality, homoscedasticity, and additivity of effects. In addition their statistical properties are nearly as good as the optimal linear models under normality assumption.

Example 9.5.2 To demonstrate the methods for combining ordered categorical data from a set of strata, we use the data of change in diagnostic confidence in Table 9.1.5, which were obtained from a clinical trial for comparing a new contrast agent with a placebo concurrent control. The stratum for this dataset is the gender of the subjects. We use integer score from 1 to 6 to denote categories from *worsened* to *excellent* in our computation. Readers can verify that the X_{EMH} is equal to 40.204 with a p -value less than 10^{-7} . Therefore at the 5% level of significance, there is a statistically significant location shift in the distribution of subjects with respect to change in diagnosis between the new contrast agent and the placebo. Table 9.5.4 provides some results for the MH estimate of the average treatment difference and its estimated large-sample variance. From Table 9.5.4,

$$\begin{aligned}\hat{\theta}_{\text{MH}} &= \frac{33.1096 + 31.3973}{(2)(18.2466)} \\ &= 1.7676\end{aligned}$$

and

$$\begin{aligned}\hat{v}(\hat{\theta}_{\text{MH}}) &= \frac{37.8475 + 37.7230}{(2 * 18.2466)^2} \\ &= 0.05674.\end{aligned}$$

Note that X_{EMH} can be easily obtained by using PROC FREQ of SAS® with the option statement CMH. One can also perform computation of $\hat{\theta}_{\text{MH}}$ and $\hat{v}(\hat{\theta}_{\text{MH}})$ in DATA statement of SAS®.

9.6 MODEL-BASED METHODS

In the previous sections we introduced several statistical methods for analysis of categorical data. These methods only require random assignment of subjects to treatments. As a result, valid statistical inference can always be obtained from these methods for randomized clinical trials. However, since these methods are randomization-based methods that are for hypothesis testing rather than estimation, they cannot describe the relationship between the categorical response and covariates such as demographical variables, patient baseline characteristics, or stratification factors. In this section we will introduce model-based procedures as alternatives to the randomization methods for analysis of categorical data.

One of the most commonly employed model-based methods for the analysis of categorical data is logistic regression (Agresti, 2002; Hosmer and Lemeshow, 2000). The method of logistic regression is useful because that (1) it can incorporate discrete covariates such as gender as well as continuous explanatory variables such as age, (2) it provides a more powerful inference for treatment effect through the reduction of variability caused by the covariates, (3) it has flexibility that allows for an adjustment of covariates, and (4) it enables the investigation of possible interactions between covariates. Thus, the method of logistic regression is usually employed to provide additional information regarding estimation, relationship between response and covariates, and homogeneity of treatment effect across different levels of covariates.

To introduce the concept of logistic regression, consider the NINDS trial. Let Y_{ghi} be the number of subjects out of n_{ghi} subjects, with time interval h to the start of treatment, who were randomly assigned to receive the i th treatment and had an early clinical improvement in study part g . Also let P_{ghi} be the corresponding probability, $g = 1, 2$, $h = 1, 2$, and $i = 1, 2$. If the subjects in each stratum formed by the complete cross-classification of part, time interval, and treatment are randomly selected independently from the corresponding targeted population, then Y_{ghi} can be described by the following product binomial distribution:

$$P\{Y_{ghi}\} = \prod_{g=1}^2 \prod_{h=1}^2 \prod_{i=1}^2 \left\{ \frac{(n_{ghi})!}{(Y_{ghi})!(n_{ghi}-Y_{ghi})!} \right\} P_{ghi}^{Y_{ghi}} (1-P_{ghi})^{(n_{ghi}-Y_{ghi})}.$$

Define the logit of P_{ghi} as

$$\text{logit}(P_{ghi}) = \ln \left\{ \frac{P_{ghi}}{1-P_{ghi}} \right\}. \quad (9.6.1)$$

The logit of P_{ghi} is therefore the logarithm of the odds of clinical improvement to no improvement for a subject with the time interval h who was randomized to receive treatment i in study part g . Logistic regression can then be used to investigate the relationship between the probability of improvement and a set of covariates by assuming that the logit of P_{ghi} is linearly related to the covariates,

$$\ln \left\{ \frac{P_{ghi}}{1-P_{ghi}} \right\} = \alpha + \mathbf{x}'_{ghi} \boldsymbol{\beta}, \quad (9.6.2)$$

where α is the intercept and $\boldsymbol{\beta}$ is a vector of unknown regression coefficients corresponding to the row vector of \mathbf{x}'_{ghi} representing study part g , time interval h , and treatment i . The expression in (9.6.2) is usually referred to as the logistic regression because the probability of improvement can be expressed as a function of explanatory variables by the logistic distribution as

$$P_{ghi}(\alpha + \mathbf{x}'_{ghi} \boldsymbol{\beta}) = \{1 + \exp[-(\alpha + \mathbf{x}'_{ghi} \boldsymbol{\beta})]\}^{-1}. \quad (9.6.3)$$

Although the logistic regression assumes a linear relationship between the logit of P_{ghi} and covariates, the equation is in fact a nonlinear function. To obtain estimates of the intercept α and regression coefficients β , the method of maximum likelihood is usually considered. The maximum likelihood estimates of α and β can be obtained by first substituting the probability of clinical improvement P_{ghi} in (9.6.1) by the right-hand side of (9.6.3). Then, differentiating the logarithm of the resulting equation with respect to α and β , we can solve the resulting normal equations for α and β . It should be noted that in general, there exist no closed forms for α and β . Thus a variety of methods for numerical optimization are usually applied to find the maximum likelihood estimators (MLE) of α and β . These methods usually involve the technique of iterative reweighted least squares (IRLS) such as the Newton-Raphson algorithm or the method of scoring which is available in most commercial statistical computer software packages such as SAS, BMDP, and GLIM. For further details, see Cox and Snell (1989), Agresti (2002), and Hosmer and Lemeshow (2000).

Let a and b denote the MLEs for α and β . Basically a and b are consistent estimates for α and β , the unknown intercept and regression coefficients, in the sense that a and b will be very close to α and β within a negligible distance when the sample size is sufficiently large. The joint distribution of a and b can be approximated by a multivariate normal distribution with mean vector $(\alpha, \beta)'$ and covariance matrix V . A consistent estimate of covariance matrix of a and b can be obtained as the following inverse of the observed information matrix given by Koch and Edwards (1988):

$$V(\alpha, \beta) = \left\{ \sum_{g=1}^2 \sum_{h=1}^2 \sum_{i=1}^2 n_{ghi} P_{ghi} (1 - P_{ghi}) (1, \mathbf{x}_{ghi})(1, \mathbf{x}_{ghi})' \right\}^{-1}. \quad (9.6.4)$$

After the MLEs a and b are obtained, the predicted probability of clinical improvement, p_{ghi} , can be obtained by replacing the unknown parameters α and β in (9.6.3) by their MLEs a and b . Hence the number of subjects with an early clinical improvement predicted by model (9.6.2) for the stratum formed by the combination of study part g , time interval h , and treatment i is given by

$$m_{ghi} = n_{ghi} p_{ghi}. \quad (9.6.5)$$

Koch and Edwards (1988) suggest using goodness-of-fit for model (9.6.2) by way of the Pearson's test or the log-likelihood chi-square test, which are given by, respectively,

$$\chi_P^2 = \sum_{g=1}^2 \sum_{h=1}^2 \sum_{i=1}^2 \left\{ \frac{(y_{ghi} - m_{ghi})^2}{m_{ghi}} + \frac{(n_{ghi} - y_{ghi})^2}{n_{ghi} - m_{ghi}} \right\}, \quad (9.6.6)$$

and

$$\chi_L^2 = \sum_{g=1}^2 \sum_{h=1}^2 \sum_{i=1}^2 2 \left\{ y_{ghi} \ln \left(\frac{y_{ghi}}{m_{ghi}} \right) + (n_{ghi} - y_{ghi}) \ln \left(\frac{n_{ghi} - y_{ghi}}{n_{ghi} - m_{ghi}} \right) \right\}, \quad (9.6.7)$$

where $y_{ghi} \ln(y_{ghi}/m_{ghi})$ and $(n_{ghi} - y_{ghi}) \ln((n_{ghi} - y_{ghi})/(n_{ghi} - m_{ghi}))$ are defined to be 0 if $y_{ghi} = 0$ or $n_{ghi} - y_{ghi} = 0$. We conclude that a significant lack-of-fit exists for the assumed model at the α level of significance if χ_P^2 or χ_L^2 is greater than the upper α th quantile of a central chi-square distribution with $GHI - (s+1)$ degrees of freedom, where G , H , and I are the number of the level for each of the three covariates and s is the number of independent explanatory variables in the model. For the data of clinical improvement in Table 9.1.2, $G = H = I = 2$ because there are two study parts, two treatments (rt-PA and the placebo),

and two time intervals from the onset of stroke to the start of treatment (0–90 and 91–190 minutes).

Logistic regression is a typical example that relates the random occurrence of clinical endpoints to the systematic components formed by a set of covariates through a link function (which is referred to as logit link). Other useful link functions include (1) the identity link and log link for the Poisson variable, (2) the probit link based on the normality assumption, and (3) the complementary log-link from the extreme-value distribution. Statistical inference based on other link functions can be similarly derived. For details regarding other link functions and their applications in clinical trials, see McCullagh and Nelder (1989).

Example 9.6.1 To illustrate the application of logistic regression, we again consider the data from the NINDS trial given in Table 9.1.2. In this trial, there are two study parts, 1 ($g = 1$) and 2 ($g = 2$), two time intervals for the onset of stroke to the start of treatment, 0 to 90 ($h = 1$) and 91 to 180 minutes ($h = 2$), and 2 treatments for this trial, placebo ($i = 1$) and rt-PA ($i = 2$). Thus, there are eight logits. Based on these logits, we can fit a main effects model with terms of *part*, *time interval*, and *treatment*. Since each covariate has two levels, for simplicity, the lower level and higher level are coded as 0 and 1, respectively, to obtain the design matrix of the main effects model. Table 9.6.1 gives the design matrix for the main effects model. Table 9.6.2 summarizes the estimates of intercept and regression coefficients, their standard errors, and the corresponding two-tailed p -values. In addition to the intercept, Table 9.6.2 gives the contribution of these three covariates. The chi-squares by log-likelihood method and the scoring method are 12.054 and 11.987, respectively. The corresponding p -values calculated based on a chi-square distribution with three degrees of freedom are given by 0.0072 and 0.0074, respectively. The results indicate that there is a significant contribution from the three main effects. Suppose that we are also interested in exploring the potential interaction and homogeneity of the treatment effect across study parts and in exploring the time to the start of treatment. Then we can include additional variables: *part*time* interaction, *part*treatment* interaction, and *time*treatment* interaction. These effects can be characterized by the product of the corresponding columns in the design matrix of the main effects model as given in Table 9.6.1. Note that the last column of Table 9.6.2 provides estimates, their standard errors, and p -values for the model

Table 9.6.1 Design Matrix for Main Effects in the Data Set of the NINDS Trial in Table 9.1.2

Part	Row Time		Column			
	Interval	Treatment	Intercept	Part	Time Interval	Treatment
1	0–90	Placebo	1	0	0	0
1	0–90	rt-PA	1	0	0	1
1	91–180	Placebo	1	0	1	0
1	91–180	rt-PA	1	0	1	1
2	0–90	Placebo	1	1	0	0
2	0–90	rt-PA	1	1	0	1
2	91–180	Placebo	1	1	1	0
2	91–180	rt-PA	1	1	1	1

Table 9.6.2 Regression Parameters and Their Estimated Standard Errors for the Data Set of the NINDS Trial in Table 9.1.2

Parameter	Statistic	Model	
		Main Effect Only	With Interaction
Intercept	Estimate (<i>se</i>)	-0.2193 (0.1682)	-0.3459 (0.2282)
	<i>p</i> -value	0.1921	0.1295
Part	Estimate (<i>se</i>)	0.0302 (0.1637)	0.0488 (0.2893)
	<i>p</i> -value	0.8538	0.8659
Time Interval	Estimate (<i>se</i>)	-0.4581 (0.1633)	-0.2085 (0.2905)
	<i>p</i> -value	0.0050	0.4730
Treatment	Estimate (<i>se</i>)	0.3137 (0.1633)	0.5337 (0.2922)
	<i>p</i> -value	0.0548	0.0678
Part*Time interval	Estimate (<i>se</i>)		-0.0454 (0.3280)
	<i>p</i> -value		0.8900
Part*Treatment	Estimate (<i>se</i>)		0.0054 (0.3280)
	<i>p</i> -value		0.9800
Time interval*Treatment	Estimate (<i>se</i>)		-0.4435 (0.3273)
	<i>p</i> -value		0.1754
<i>Model Chi-Square for Covariates</i>			
<i>Loglikelihood</i>			
<i>df</i>		3	6
Chi-square		12.054	13.917
<i>p</i> -value		0.0072	0.0306
<i>Score</i>			
<i>df</i>		3	6
Chi-square		11.987	13.980
<i>p</i> -value		0.0074	0.0299

with two-factor interactions. The contributions of the six explanatory variables assessed by both the log-likelihood and scoring methods are statistically significant at the 5% level of significance. However, the additional contribution of the three two-factor interactions based on the log-likelihood method and the scoring method are given by 1.863 and 1.993, respectively, which are not statistically significant at the 5% level of significance. For the goodness-of-fit of the main effects model, χ^2_P and χ^2_L test statistics gave 5.273 and 5.295 with corresponding *p*-values of 0.260 and 0.258, respectively. In conclusion, the main effects model is an adequate model for describing the relationship between the probability of clinical improvement and study parts, time to the start of treatment, and treatment.

The results in Table 9.6.2 indicate that the probability of an early clinical improvement for a subject with time of onset of stroke to treatment within 90 minutes is statistically significantly larger than that for a subject with time to treatment between 90 and 180 minutes. However, the treatment effect is marginally significant with a *p*-value of 0.0548. An analysis by strata reveals that subjects with the time to treatment within 90 minutes in the study part 2 had a significant treatment effect. In addition to the observed frequency and probability of clinical improvement, Table 9.6.3 displays the predicted frequency and probability by

Table 9.6.3 Observed and Predicted Frequencies and Probabilities of Clinical Improvement for the Data Set of the NINDS Trial in Table 9.1.2

Part	Time Interval	Treatment	Sample Size	Frequency		Probability (%)	
				Observed	Predicted	Observed	Predicted
1	0–90	Placebo	68	31	30.29	45.6% (6.04%)	44.5% (4.15%)
	0–90	rt-PA	71	36	37.17	50.7% (5.93%)	52.4% (4.13%)
	91–180	Placebo	79	26	26.61	32.9% (5.29%)	33.7% (3.70%)
1	91–180	rt-PA	73	31	29.93	42.5% (5.79%)	41.0% (4.02%)
	0–90	Placebo	77	30	34.87	39.0% (5.56%)	45.3% (4.02%)
	0–90	rt-PA	86	51	45.67	59.3% (5.30%)	53.1% (4.02%)
2	91–180	Placebo	88	35	30.24	39.8% (5.22%)	34.4% (3.64%)
	91–180	rt-PA	82	29	34.22	35.4% (5.28%)	41.7% (3.90%)

the fitted main effects model and their standard errors. Note that the standard error of the observed probability is larger than that of the predicted probability because of the reduction of variability induced by the covariates.

The estimated logits and their standard errors and odds of improvement with the corresponding 95% confidence intervals are given in Table 9.6.4. For the situation where the time from the onset of stroke to the start of treatment is greater than 90 minutes, the upper limit of the 95% confidence interval is less than 1. As a result the odds of clinical improvement is statistically significantly less than 1 at the 5% level of significance regardless of which treatment a subject is given for both parts of the study. When the time of onset to the start of treatment is within 90 minutes, the odds of clinical improvement is not statistically significant, since its 95% confidence interval includes 1. Note that the difference in estimated logits between rt-PA and the placebo is the same across the four (part-by-time interval) strata. This difference is exactly the same as the estimate for the regression coefficient corresponding to the treatment given by 0.3137. The common odds ratio of clinical improvement for a subject receiving rt-PA to a subject given the placebo is estimated as $\exp(0.3137) = 1.368$. The corresponding 95% confidence interval can be calculated as $\exp\{0.3137 \pm (1.96)(0.1633)\}$ which gives (0.9937, 1.8847). These results are consistent with those obtained by the Mantel-Haenszel method discussed in Example 9.5.1.

The ordinal categorical data such as the AUA-7 symptom score (Table 9.1.4), change in diagnostic confidence (Table 9.1.5), and status score (Table 9.1.6) have more than two categories in ascending order. One of the possible models for investigating the relationship between probabilities and covariates is the proportional odds model based on the cumulative probabilities (Agresti, 2002). For simplicity, consider a trial conducted for comparing a test drug with a placebo.

Let P_{ij} be the probability of observing the j th possible outcome (in a total of J categories of clinical responses) for subjects receiving the i th treatment, $i = 1, 2$. The cumulative logit is defined as the logarithm of the odds of the cumulative probabilities below the j th category which is given by

$$\text{logit}(P_{i1} + \dots + P_{ij}) = \ln \left\{ \frac{P_{i1} + \dots + P_{ij}}{1 - (P_{i1} + \dots + P_{ij})} \right\}, \quad (9.6.8)$$

Table 9.6.4 Estimated Logits and Odds of Clinical Improvement for the Data Set of the NINDS Trial in Table 9.1.2

Part	Time Interval	Treatment	Sample Size	Logit Estimate	Standard Error	Odds of Improvement	
						Estimate	95% CI
1	0–90	Placebo	68	-0.2194	0.1682	0.803	(0.578, 1.117)
1	0–90	rt-PA	71	0.0943	0.1654	1.099	(0.795, 1.520)
1	91–180	Placebo	79	-0.6775	0.1656	0.508	(0.367, 0.703)
1	91–180	rt-PA	73	-0.3638	0.1661	0.695	(0.502, 0.963)
2	0–90	Placebo	77	-0.1892	0.1624	0.828	(0.602, 1.138)
2	0–90	rt-PA	86	0.1245	0.1581	1.133	(0.831, 1.544)
2	91–180	Placebo	88	-0.6473	0.1613	0.523	(0.362, 0.718)
2	91–180	rt-PA	82	-0.3337	0.1606	0.716	(0.523, 0.981)

where $j = 1, \dots, J - 1$. A reasonable model is to assume a linear regression with different intercepts for each of the $J - 1$ cumulative logits and a common slope as follows:

$$\text{logit}(P_{i1} + \dots + P_{ij}) = \alpha_j + \beta x, \quad (9.6.9)$$

where $x = 1$ if the treatment is the test drug and $x = 0$ if the treatment is the placebo and $j = 1, \dots, J - 1$. This model assumes that the treatment effect on the odds of response below the j th category is the same for all j . In other words,

$$\begin{aligned} & \text{logit}(P_{11} + \dots + P_{1j}) - \text{logit}(P_{01} + \dots + P_{0j}) \\ &= \ln \left\{ \frac{P(Y \leq j|i=1)/P(Y > j|i=1)}{P(Y \leq j|i=0)/P(Y > j|i=0)} \right\} \\ &= \alpha_j + \beta - \alpha_j \\ &= \beta. \end{aligned} \quad (9.6.10)$$

The first expression within the logarithm on the left-hand side of equation (9.6.10) is usually referred to as the cumulative odds ratio. The model in (9.6.9) assumes that the logarithm of the cumulative odds is proportional to the difference between the values of the explanatory variables for all j . Because of this property, model (9.6.9) is also called the *proportional odds model*. Within the same treatment, we have

$$\text{logit}(P_{i1} + \dots + P_{ij}) - \text{logit}(P_{i1} + \dots + P_{i(j-1)}) = \alpha_j - \alpha_{j-1}, \quad (9.6.11)$$

where $j = 1, \dots, J - 1$. Since we assume a common slope for each cumulative logit, they are parallel to each other by an amount $\alpha_j - \alpha_{j-1}$, $j = 1, \dots, J - 1$. As a result it is important to check the parallel lines assumption before the proportional odds model can be applied.

Suppose that there are s covariates. One can fit a full model to the cumulative logits without assuming common slopes. The number of parameters for the full model is $s(J - 1)$, so a proportional odds model has a total of $J - 1 + s$ parameters. Consequently, the chi-square score statistic for the parallel lines assumption approximately follows a chi-square distribution with $s(J - 1) - (J - 1 + s) = s(J - 2)$ degrees of freedom. Consequently, one rejects the assumption of the proportional odds model if the chi-square score statistic is too large.

Example 9.6.2 The data on question 4 given in Table 9.1.4, concerning the urination frequency within two hours of the last urination in the validation study comparing subjects with BPH and normal controls, are used to illustrate the application of the proportional odds model. For this dataset, $J = 6$ and $s = 1$. The observed chi-square score statistic is 1.5378 with $4 (= 6 - 2)$ degrees of freedom. Since the corresponding p -value is 0.8198, there is no evidence to suspect the validity of the parallel lines assumption. Thus the proportion odds model is applied to this dataset. Table 9.6.5 provides estimates of five intercepts, the common treatment effects, and their corresponding standard errors. The results are summarized in Table 9.6.5. From Table 9.6.5, it can be seen that the treatment effect is estimated by 1.3192 with a standard error of 0.3364 which is statistically, significantly different from zero. The common cumulative ratio for observing a response below the j th category from a normal control to the BHP subjects is 3.7404 with the corresponding confidence interval of (1.9345, 7.2322).

Table 9.6.5 Estimates of Regression Parameters and Their Estimated Standard Errors for Question 4—Frequency No. 1 in AUA Symptom Index

Parameter	Statistics	Value
Intercept 1	Estimate	-3.5099
	<i>se</i>	0.5797
	<i>p</i> -value	<0.0001
Intercept 2	Estimate	-1.8712
	<i>se</i>	0.5115
	<i>p</i> -value	0.0003
Intercept 3	Estimate	-1.0113
	<i>se</i>	0.4920
	<i>p</i> -value	0.0398
Intercept 4	Estimate	-0.0018
	<i>se</i>	0.5000
	<i>p</i> -value	0.9971
Intercept 5	Estimate	0.8619
	<i>se</i>	0.5485
	<i>p</i> -value	0.1161
Treatment	Estimate	1.3192
	<i>se</i>	0.3364
	<i>p</i> -value	<0.0001

9.7 REPEATED CATEGORICAL DATA

As indicated in the previous chapter, repeated measures are often obtained at various time points after the baseline has been established in order to evaluate the efficacy and safety of a drug under study. For drug efficacy, repeated categorical data provide valuable information regarding the time-response profile which is often used to determine (1) how long the drug needs to be administered before its clinically meaningful effectiveness is achieved and (2) how long the effectiveness can be maintained by the treatment. For safety, repeated categorical measurements can also be used to characterize the time course of the occurrence of adverse events.

A commonly used approach for the evaluation of a drug is to compare the differences in changes of categorical data from the baseline between groups. For example, for data given in Table 9.2.3 which were used to illustrate the McNemar test in Section 9.2, the proportion of subjects with a favorable status in the placebo group is 45.61% at both the baseline and visit 3. As a result there is no change from the baseline at visit 3 in the proportion of favorable status for subjects receiving the placebo. On the other hand, for those subjects receiving the test drug, the proportions of favorable status are 42.59% and 72.22% at the baseline and visit 3, respectively. This indicates a 29.63% increase in proportion with favorable status for the test group. As a result it is of interest to test whether the observed 29.63% difference in the change in proportion with favorable status from the baseline between the test drug and the placebo observed at visit 3 is statistically significantly different from zero. For this purpose, we could consider the changes in marginal probabilities between visit 3 and the baseline as the response variables for comparison between treatments. This problem is referred to as the two-sample McNemar test by Feuer and Kessler (1989).

Let θ_g be the change in proportion from the baseline at a certain post-treatment visit for group g :

$$\begin{aligned}\theta_g &= P_{g,1} - P_{g1.} \\ &= P_{g01} - P_{g10},\end{aligned}\quad (9.7.1)$$

where $g = T, R$. The parameter of interest is then the difference in change of proportion from the baseline between treatments. That is,

$$\delta = \theta_T - \theta_R. \quad (9.7.2)$$

Therefore the hypotheses for the two-sample McNemar test can be formulated as follows:

$$\begin{aligned}H_0: \delta &= 0, \\ \text{vs. } H_a: \delta &\neq 0.\end{aligned}\quad (9.7.3)$$

Similar to the one-sample McNemar test, the maximum likelihood estimator (MLE) for θ_g can be obtained by replacing the unknown parameters in (9.7.1) by their corresponding observed sample proportions as

$$\hat{\theta}_g = p_{g01} - p_{g10}.$$

An estimate of the large-sample variance of $\hat{\theta}_g$ is then given by

$$\hat{v}(\hat{\theta}_g) = \frac{(p_{g01} + p_{g10}) - (p_{g01} - p_{g10})^2}{n_g}, \quad (9.7.4)$$

where $g = T, R$. Consequently, the MLE for δ and its large-sample variance are given by, respectively,

$$\hat{\delta} = \hat{\theta}_T - \hat{\theta}_R \quad (9.7.5)$$

and

$$v(\hat{\delta}) = v(\hat{\theta}_T) + v(\hat{\theta}_R). \quad (9.7.6)$$

Feuer and Kessler (1989) suggest using

$$X = \frac{\hat{\delta}^2}{v(\hat{\delta})} \quad (9.7.7)$$

as the test statistic for hypotheses (9.7.3) which approximately follow a chi-square distribution with one degree of freedom. Therefore, the null hypothesis of (9.7.3) is rejected at the α th level of significance if $X > \chi^2(\alpha, 1)$. The corresponding large-sample $(1 - \alpha)$ 100% confidence interval for δ is given as

$$\hat{\delta} \pm Z(\alpha/2) \sqrt{v(\hat{\delta})}. \quad (9.7.8)$$

As indicated before, many clinical trials are conducted at different centers with the same study protocol. The Mantel-Haenszel type of technique can also be applied to combine the

results from different centers for the differences in changes from the baseline in marginal proportions between treatments. Denote $\hat{\delta}_h$ as the observed difference in change in proportions of favorable status from baseline between treatments for the h th center at a particular post-treatment visit. Also let $v(\hat{\delta}_h)$ be its large-sample variance, where $h = 1, \dots, H$. Then a test statistic for testing hypotheses (9.7.3) is the ratio of the square of the sum of $\hat{\delta}_h$ to the sum of their variances:

$$X_C = \frac{\left[\sum_{h=1}^H \hat{\delta}_h \right]^2}{\sum_{h=1}^H v(\hat{\delta}_h)}. \quad (9.7.9)$$

When the total sample size is sufficiently large, the distribution of X_C approximately follows a chi-square distribution with one degree of freedom. As a result, we can reject the null hypothesis of no difference in change from the baseline in proportions between treatments at the α th level of significance if $X_C > \chi^2(\alpha, 1)$.

As described in Example 9.2.2, the binary data for the proportion with favorable status (Table 9.2.3) are obtained by re-grouping the five categories of the original data (Table 9.1.6). We can also directly compare the transition in marginal proportions between treatments without combining classes into a binary response. The two-sample McNemar test for comparing change in proportion from the baseline between groups can be extended to the categorical data with more than two classes. We first consider the one-sample case. Suppose that a categorical datum has a total of r categories. Let P_{ij} be the proportion of subjects who had a response in the i th category at the baseline and a response in the j th category at a post-treatment visit, $i, j = 1, \dots, r$. It follows that

$$P_{\cdot i} = \sum_{j=1}^r P_{ij}$$

and

$$P_{i \cdot} = \sum_{j=1}^r P_{ij}$$

are the marginal proportions of the number of subjects with a response in the i th category at the baseline and that with a response in the j th category at a post-treatment visit, respectively. Within the same group, one can test whether the marginal proportions are the same for both the baseline and the visit. The hypotheses of interest can be expressed as follows:

$$\begin{aligned} H_0: P_{\cdot i} &= P_{i \cdot} \quad \text{for all } i, \\ \text{vs.} \quad H_a: P_{\cdot i} &\neq P_{i \cdot} \quad \text{for at least one } i, i = 1, \dots, r. \end{aligned} \quad (9.7.10)$$

Transition in the proportions from baseline to visit can be summarized by an $(r - 1) \times 1$ vector $\mathbf{d} = (d_1, d_2, \dots, d_{r-1})'$, where $d_i = p_{i \cdot} - p_{\cdot i}$ is the difference in sample proportions between the baseline and a visit for the i th category $i = 1, \dots, r - 1$. The large-sample covariance matrix for d_i can then be estimated by

$$\frac{\mathbf{W}}{n} = \frac{\{w_{ij}\}}{n},$$

where

$$w_{ij} = \begin{cases} (p_{i\cdot} + p_{\cdot i}) - 2p_{ii} - (p_{i\cdot} - p_{\cdot i})^2 & \text{if } i = j, \\ -(p_{ij} + p_{ji}) - (p_{i\cdot} - p_{\cdot i})(p_{j\cdot} - p_{\cdot j}) & \text{if } i \neq j, \end{cases} \quad (9.7.11)$$

where $1 \leq i, j \leq r - 1$. A test statistic proposed by Bhapkar (1966) for the homogeneity of marginal proportions between the baseline and the visit is given by

$$Q = n\mathbf{d}'\mathbf{W}^{-1}\mathbf{d},$$

which follows a central chi-square distribution with $r - 1$ degrees of freedom. Therefore the null hypothesis of (9.7.10) is rejected at the α th level of significance if $Q > \chi^2(\alpha, r - 1)$, where $\chi^2(\alpha, r - 1)$ is the upper α th quantile of a chi-square distribution with $r - 1$ degrees of freedom. The covariance under the null hypothesis, $P_{i\cdot} = P_{\cdot i}$, for all i , is given as

$$w_{ij}^0 = \begin{cases} (p_{i\cdot} + p_{\cdot i}) - 2p_{ii}, \\ -(p_{ij} + p_{ji}), \end{cases} \quad (9.7.13)$$

where $1 \leq i, j \leq r - 1$. Replacing \mathbf{W} in the Bhapkar's Q statistic with the null covariance matrix \mathbf{W}^0 , the resulting test statistic is the one proposed by Stuart (1955) which has the same asymptotic distribution as Bhapkar's Q .

Let $\boldsymbol{\theta}_h$ represent the vector consisting of $r - 1$ changes in marginal proportions from the baseline, where

$$\boldsymbol{\theta}_h = (P_{h1\cdot} - P_{1\cdot 1}, \dots, P_{h(r-1)\cdot} - P_{(r-1)\cdot 1}), \quad h = T, R. \quad (9.7.14)$$

Then $\boldsymbol{\delta} = \boldsymbol{\theta}_T - \boldsymbol{\theta}_R$ is the parameter of interest, which is the difference in changes in proportions from the baseline between treatments. The hypothesis on whether there is a difference in the change in marginal proportions from the baseline can then be formulated as

$$\begin{aligned} H_0: \boldsymbol{\delta} &= 0, \\ \text{vs. } H_a: \boldsymbol{\delta} &\neq 0. \end{aligned} \quad (9.7.15)$$

Let \mathbf{d}_h be the vector of the differences obtained from the observed sample proportions for the h th group based on sample size n_h with its estimated non-null sample covariance matrix \mathbf{W}_h . Then the Bhapkar's Q statistic for testing hypotheses of (9.7.15) is given by

$$Q = (\mathbf{d}_T - \mathbf{d}_R)' \left[\frac{\mathbf{W}_T}{n_T} + \frac{\mathbf{W}_R}{n_R} \right]^{-1} (\mathbf{d}_T - \mathbf{d}_R). \quad (9.7.16)$$

This statistic approximately follows a chi-square distribution with $r - 1$ degrees of freedom. As a result we reject the null hypothesis at the α th level of significance if $Q > \chi^2(\alpha, r - 1)$.

Note that the methods for analysis of repeated categorical data discussed in this section are based on a change from the baseline in marginal proportions. Other functions of the marginal probabilities such as cumulative logits described in (9.6.8) can be used as response variables for the characterization of change from the baseline disease status.

Repeated categorical data are measurements of the same categorical end-point on the same individuals over the course of a trial. For the same subject repeated categorical data are correlated to each other. Analysis of repeated categorical data by treating the repeated data as independent observations will result in some undesirable statistical deficiencies. First of all, the estimates of treatment effects may not be robust to the selection bias for nonrandomized trials. Second, statistical inference for an average response is not efficient, although sometimes it is valid. Finally, estimation for variability is inconsistent. Alternatively, as discussed in the previous chapter, the technique of GEE developed by Liang and Zeger (1986) and Zeger et al. (1988) can be applied for analysis of repeated categorical data using either the marginal or transition models. Miller et al. (1993) propose a procedure for the analysis of repeated categorical data with more than two classes using the GEE in connection with weighted least squares. Lipsitz et al. (1994) report a performance of the GEE in practical circumstances. Agresti (1983, 1984, 1989, 2002) provides a comprehensive review of models and methods for analysis of repeated ordered categorical response data. Zeger and Liang (1992) give an overview of methods for the analysis of longitudinal data. Recent developments on repeated categorical data can be found in Crowder and Hand (1990), Lindsey (1993), Diggle, Liang, and Zeger (1994), and Hand and Crowder (1996).

Example 9.7.1 Again we use the binary response of favorable status to illustrate the two-sample McNemar test. Table 9.7.1 presents the results of a one-sample McNemar test using the non-null variance given in (9.7.4). The results are consistent with those given in Table 9.2.5. The MLE for the proportional difference in change of favorable status from the baseline at visit 3 is 29.63%, with an estimated standard error of 0.0972. As a result the large-sample 95% confidence interval calculated according to (9.7.8) is (10.59%, 48.68%), which does not contain zero. Hence we reject the null hypothesis at the 5% level of significance that the change in proportions at visit 3 from the baseline between treatments is the same. The conclusion is that significantly more subjects receiving the test drug than those given the placebo changed from their unfavorable status at the baseline to the favorable status at visit 3.

Next we use the status score data in the original categories for illustration of Bhapkar's procedures. Since the data are sparse in the *terrible* response category, we will combine the *terrible* and *poor* response categories. Table 9.7.2 gives the changes in marginal proportion at visit 3 from the baseline and their differences between treatments. The one-samples of Bhapkar's Q statistics are 21.82 and 8.93 for the test drug and the placebo with p -value < 0.00001 and p -value = 0.0302, respectively. The results indicate that at visit 3 there is a statistically significant change in marginal proportions from the baseline for both treatments. The two-sample of Bhapkar's Q is given by 12.05 with a p -value of 0.0072. As a result the null hypothesis of no difference in change of marginal proportions between

Table 9.7.1 Summary of Differences in Favorable Status for Data Set Provided in Table 9.2.3

Treatment	N	Difference in Proportion	Standard Error	Chi-square	p -Value
Test drug	54	0.2963	0.0674	19.31	<0.0001
Placebo	57	0.0000	0.0702	0.000	>0.9999
Difference	111	0.2963	0.0972	9.27	0.0023

Table 9.7.2 Summary of Marginal Changes in the Data Set of Table 9.1.6

Treatment	N	Time Point	Category				
			Terrible or Poor	Fair	Good	Excellent	
Test drug	54	Visit 3	9.26%	18.52%	25.92%	46.30%	
		Baseline	24.07%	33.33%	22.22%	20.37%	
			-14.81%	-14.81%	3.70%	25.93%	
Change							
Bhapkar $Q = 21.82$							
with 3 df							
p -value < 0.0001							
Placebo	57	Visit	29.83%	24.56%	19.30%	26.32%	
		Baseline	19.30%	35.09%	33.33%	12.28%	
			10.63%	-10.53%	-14.03%	14.04%	
Change							
Bhapkar $Q = 8.93$							
with 3 df							
p -value = 0.0302							
Difference in change							
Bhapkar $Q = 12.05$							
with 3 df							
p -value = 0.0072							

treatments is rejected at the 5% level of significance. From Table 9.7.2 more subjects on the test drug than on the placebo changed from other statuses at the baseline to the excellent rating at visit 3.

9.8 DISCUSSION

In this chapter, we provided a concise review and introduction to the concept and several useful statistical methods commonly employed for the analysis of categorical data. However, we noted that some methods appearing in the literature are not discussed. These methods include (1) methods that count data using a Poisson distribution and (2) certain model-based methods such as the loglinear model. For more detail on the analysis of categorical data and related topics, see Agresti (2002).

The statistical tests discussed in this chapter are mainly for testing hypothesis regarding the existence of treatment effect. These methods are often misused to establish therapeutic equivalence between a test drug and a reference drug. For testing therapeutic equivalence for the binary response, as indicated in Chapters 2 and 6, the following interval hypotheses should be considered

$$\begin{aligned} H_0: P_{11} - P_{21} &\geq U \quad \text{or} \quad P_{11} - P_{21} \leq L, \\ \text{vs. } H_a: L < P_{11} - P_{21} &< U. \end{aligned}$$

We conclude that the test drug is therapeutically equivalent to the reference drug at the α level of significance if the $(1-2\alpha)100\%$ confidence interval for the difference in population proportions computed from (9.3.3) falls within the equivalence limits of (L, U) . In

clinical trials, however, the one-sided equivalence is more appealing because the primary objective is usually to verify whether the efficacy of the test drug is at least as good as that of the reference drug. Therefore we can conclude that the test drug is at least as effective as the reference drug and conclude that equivalence is at the α level of significance if the lower limit of the $(1 - 2\alpha)100\%$ confidence interval is greater than the lower equivalence L . For more detail on therapeutic equivalence, see Blackwelder (1982), Huque and Dubey (1990), Durrleman, and Simon (1990), Makuch and Johnson (1990), Farrington and Manning (1990), Chow and Liu (2000), and Jennison and Turnbull (1993). For recent developments of equivalence/noninferiority based on binary endpoints, see Chan (2003) for a comprehensive review of methods for independent samples and see Hsueh et al. (2001) and Liu et al. (2002) for paired samples.

The Bhapkar's Q for the detection of a marginal difference in the change from the baseline can be applied to categorical responses on either a nominal or an ordinal scale. However, the method does not take into account the order of categories. Agresti (1983) proposes a more powerful Mann-Whitney type statistic for testing the homogeneity of marginal proportions between the visit and the baseline when the true marginal proportions are stochastically ordered. Agresti's procedure is useful when the categorical response is ordinal and the number of ordered categories is large.

As discussed in the previous chapter, missing values or dropouts often occur in clinical trials. Little and Rubin (1987) and Little (1995) gave definitions of a hierarchy of missing value mechanisms. In general, the entire dataset in which a clinical trial is supposed to collected can be partitioned into two parts. The first part, say \mathbf{Y}_0 , concerns data observed and collected from the trial; the second part, the data, denoted by \mathbf{Y}_m , concerns the data that were supposed to be observed but were not and hence are missing. For instance, let R be a random variable that indicates whether the data of subjects are observed or missing. The missing value mechanism is said to be missing completely at random (MCAR) if R is independent of both \mathbf{Y}_0 and \mathbf{Y}_m . If R is only independent of \mathbf{Y}_m , the data are said to be missing at random (*MR*) or ignorable (Laird, 1988). The missing mechanism is said to be informative or unignorable if random variable R depends on \mathbf{Y}_m . Currently, there are no well-developed methods that can satisfactorily analyze the ordered categorical data with intermittently missing values. If the mechanism for the missing values is completely random, Lachin (1988b) indicates that inference based on the observed complete data can be treated as a subgroup analysis and hence is valid but less efficient. However, the assumption of independence between R and $(\mathbf{Y}_0, \mathbf{Y}_m)$ and between R and treatment is difficult to verify. Statistical procedures were proposed by Diggle (1989) and Ridout (1991) for testing the hypothesis of completely random dropouts, and for missing patterns of the repeated categorical data by Park and Davis (1993). For repeated categorical data, the issue of premature dropout is much more complicated than that of intermittent missing data because it almost certainly suggests that the assumption of MCAR is implausible. Despite the research effort for the informative withdrawals by prominent statisticians such as Fitzmaurice et al. (1995) and Diggle and Kenward (1994), further work is needed on repeated categorical data.

10

CENSORED DATA AND INTERIM ANALYSIS

10.1 INTRODUCTION

Statistical concepts and methods for analyzing continuous and categorical endpoints in clinical trials were reviewed in the previous two chapters. Clinical endpoints for assessment of efficacy and safety of a promising therapy usually include occurrence of some predefined events such as death, the response to a new chemotherapy in treatment of some advanced cancer, the eradication of an infection caused by a certain microorganism (e.g., *Helicobacter pylori* for gastric ulcers), serious adverse events (e.g., neutropenia), or the elevation of aspartate transaminase three times over the upper limit of the normal range. For these events, the primary parameter of interest is usually time to the occurrence of such an event. Another parameters of interest is the median survival time. The median survival time is defined as the time at which 50% of the subjects still survive. Subjects are recruited into the trial at different calendar time point. Note that the predefined event may not be observed on the subjects who complete the scheduled duration of treatment and follow-up. On the other hand, some subjects may withdraw prematurely without observing any occurrences of the event before the end of the study. These individuals are said to be lost to follow-up. As a result we do not have any information on these subjects with respect to the event. The only information we have is that the predefined event did not occur at these subjects in their last visit (either at the end of study or at the time they dropped out from the study). The time to the occurrence of the event therefore is not known for these subjects. We refer an endpoint of this kind to as a censored endpoint. As an example, consider the most common opportunistic infection in patients with advanced stages of AIDS, which is the disseminated infection with *Mycobacterium avium* complex. Pierce et al. (1996) conducted a randomized, two-group parallel, placebo-controlled double-blind study to assess the prophylactic effect

of clarithromycin 500 mg twice daily in prevention of *M. avium* complex infection and improvement of survival in patients with advanced AIDS. The entrance criteria for this study included blood cultures that were negative to *M. avium*, a Karnofsky performance score of 50 or higher, a CD4 cell count of 100 or fewer per cubic millimeter. The primary efficacy endpoints consist of the time to the detection of *M. avium*, which is defined as the time interval between randomization and the first positive blood culture, and survival, which is defined as the time interval from randomization to death from any causes. The data from the patients without infection were censored at the time of the last negative culture for the time to the detection. The survival data were censored at the time of the last contact of the alive patients. The sample size for this study is 300 patients per group to provide at least an 80% power for detection of at least a 67% reduction in the incidence of infection of *M. avium* complex for the clarithromycin group as compared with the placebo group. This trial also specified an interim analysis at the time when the first 300 patients have completed one year of therapy or when 50 patients developed *M. avium* infection, whichever came first. The study was in fact terminated after the first interim analysis because compelling evidence of reduction in incidence of *M. avium* complex provided by clarithromycin.

The Economist (2002) reports that only 1 out of 10,000 molecules screened in a given program can make it from the stage of drug discovery to marketing launch. The process of drug development and research is lengthy and on average takes 12 years. The cost increased from 125 million in US dollars in 1963–1975 to 450 million US dollars in 1991–1995 and to 800 million US dollars in 2002. In addition, the pharmaceutical industry spent 44 billion US dollars in 2001 for research and development. According to The Economist (2002), about a decade ago, testing a drug on 1,000 patients was enough to prove its safety and efficacy. Now, regulatory authorities require a database of 4,000 patients from clinical trials conducted in different development stages. Most of the budget for research and development goes to clinical trials in clinical development. About 60% of drugs, however, fail during the clinical development stage and the success rate of drugs from expensive phase III confirmatory trials to market has fallen by almost 30%. Since drug research and development is not only time-consuming but also costly and risky, it is helpful to determine whether the efficacy and safety of a new pharmaceutical entity is promising and compelling at early stage of clinical development, be that positive or negative. The information can be used to recommend early termination of the drug's development. This is especially important for trials with long-term follow-up or intended for evaluation of life-threatening or severely debilitating diseases. As a result interim analyses have become popular in clinical development. The purpose is to provide early information regarding the test drug in a more cost-effective way. If the reason for early termination is lack of efficacy or the presence of unexpected or untoward adverse experiences, then the remaining resource may be re-allocated to develop other promising drugs. Therefore interim analyses to provide statistical evidence for early termination and data monitoring have become an indispensable norm in the management of clinical trials. However, it should be noted that interim analysis and data monitoring are two different areas. They have their own functions and procedures, although they are related to each other.

Analysis of censored data and the application of interim analysis for early termination have become common practice for clinical trials. In this chapter, we will first discuss different types of censoring. Some background regarding the survival function and the use of the Kaplan-Meier's method (Kaplan and Meier, 1958) for estimation of proportions based on censored data are given in Section 10.2. The logrank statistic and Gehan's generalization of Wilcoxon rank sum statistic for the censored data (Gehan, 1965a, 1965b) are outlined in Section 10.3. Section 10.4 describes the semiparametric proportional hazard

model with covariates as proposed by Cox (1972). The concept of calendar and information time are illustrated in Section 10.5. Different methods for interim analysis are covered in Section 10.6. These methods include the group sequential methods proposed by Pocock (1977) and O'Brien and Fleming (1979), repeated confidence intervals suggested by Jennison and Turnbull (1989), the method of alpha spending function by DeMets and Lan (1994), and *B*-values by Lan and Wittes (1988). Some final remarks and a discussion will be provided in the last section. Whenever possible, numerical examples using published data from real trials will be used to illustrate the methods presented in this chapter.

10.2 ESTIMATION OF THE SURVIVAL FUNCTION

In conducting clinical trials, it is almost impossible and impractical to enroll all the patients into a study at the same time. Patients are usually enrolled into the trial at different time points during the study. At the end of the study or at the time of interim analysis, the event of interest may occur in some patients during the study. As a result, their times to the event are actually observed. However, for the patients who completed the study or withdrew prematurely from the study without the event, their times to the events may not be observed and hence are censored (right censored). Figure 10.2.1 displays in calendar time the censoring patterns for time to a predefined event of a study, starting from the initiation to its end or to the time at which an interim analysis was performed. Patient 1 enrolled at the beginning of the study and the event of interest occurred at month 10. Therefore, the time to the event is 10 months. Patient 2 also entered at the beginning of the study, but for some reason this patient prematurely withdrew from the study at month 8 without the event. Hence, the time to the event is unknown and censored at 8 months. Patient 3 completed the entire study of 16 months without an occurrence of the event. Therefore, the time to the event of patient 3 is censored at 16 months. Patient 4 was enrolled at month 2, and the event occurred at month 12. As a result, the time to the event for patient 4 is 10 months. Patient 5 entered into the trial at month 4 but lost to follow-up. Hence, the time to the event for patient 5 is censored at 8 months. The time to the event is 8 months and not censored for patient 6 who entered into

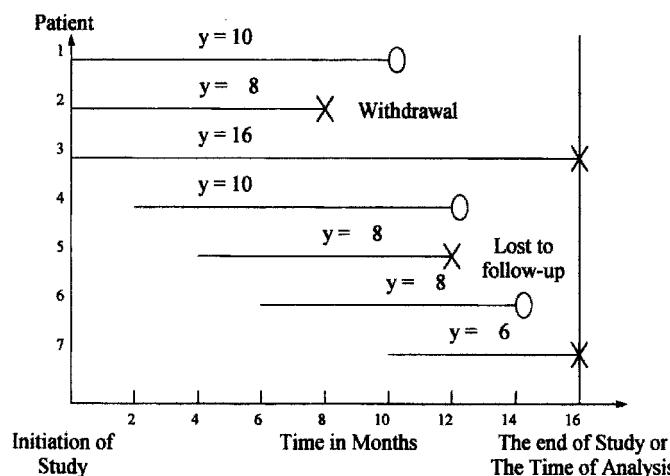


Figure 10.2.1 Censoring pattern in calendar time. ○: event; ×: censored.

the trial at month 6 and experienced the event at month 14. Patient 7 was enrolled into the study at month 10. When the study ended at month 16, no occurrences of the event were observed and hence the time to the event for patient 7 is censored at 6 months. In survival analysis, it is usually of interest to obtain statistical inference on the time from initiation of treatment such as randomization to the occurrence of the predefined event. As a result, it is suggested that the data be arranged in terms of the length of treatment before the occurrence of the event as illustrated in Figure 10.2.2. As can be seen from Figure 10.2.2, the times to the event for patient 1 and patient 4 are the same (10 months). Similarly, the times to the event for patient 2 and patient 5 are both censored at 8 months.

Fleming et al. (1980) report the results from a study conducted in patients with limited stage II or IIIA ovarian carcinoma at the Mayo Clinic which was to determine whether the grade was related to the time over which disease progressed. Table 10.2.1 displays either the observed time in months to the progression of the disease or the time censored at the last visit. The observed time is given in the second column, while the third column provides information on whether the observed time to progression was relevant or censored. The last column indicates whether a patient had low grade or well-differentiated cancer. Table 10.2.1 provides a basic layout for the presentation of censored data which must include patient identification, time to the event, censored indicator, and other covariates such as treatment assignment and disease status.

Let Y be a continuous random variable representing the time from randomization to the occurrence of a clinically meaningful event such as the time until the detection of *M. avium* or the occurrence of death for the AIDS trial as described in the previous section. Y is usually referred to as the survival time. The cumulative distribution function (cdf) of Y , denoted by $F(y)$, is defined as the probability that a subject fails before or equal to the time y , namely

$$\begin{aligned} F(y) &= P(\text{a subject fails before or equal to } y) \\ &= P(Y \leq y), \quad 0 < y < \infty. \end{aligned} \quad (10.2.1)$$

The cdf is a nondecreasing function of time y such that $F(0) = 0$ and $F(\infty) = 1$. The cdf of Y is usually employed to describe the mortality rate for evaluation of treatment in

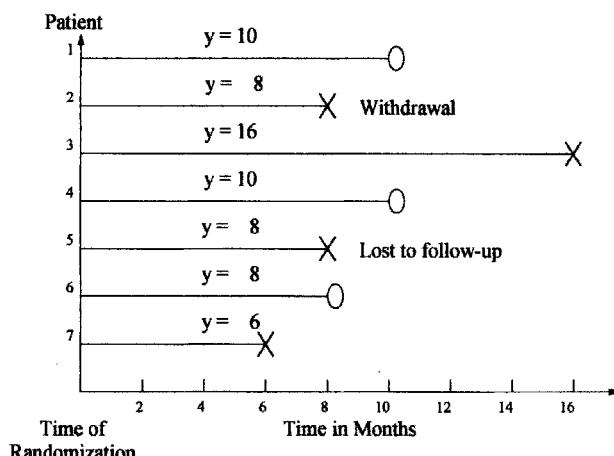


Figure 10.2.2 Censoring pattern in duration of treatment. ○: event; ×: censored.

Table 10.2.1 Time in Months to Progression of the Patients with Stage II or IIIA Ovarian Carcinoma by Low-Grade or Well-Differentiated Cancer

Patient Number	Time in Months	Censored	Cell Grade
1	0.92	No	Low
2	2.93	No	Low
3	5.76	No	Low
4	6.41	No	Low
5	10.16	No	Low
6	12.40	Yes	Low
7	12.93	Yes	Low
8	13.85	Yes	Low
9	14.70	Yes	Low
10	15.20	No	Low
11	23.32	Yes	Low
12	24.47	Yes	Low
13	25.33	Yes	Low
14	36.38	Yes	Low
15	39.67	Yes	Low
16	1.12	No	High
17	2.89	No	High
18	4.51	No	High
19	6.55	No	High
20	9.21	No	High
21	9.57	No	High
22	9.84	Yes	High
23	9.87	Yes	High
24	10.16	No	High
25	11.55	No	High
26	11.78	No	High
27	12.14	No	High
28	12.14	No	High
29	12.17	No	High
30	12.34	No	High
31	12.57	No	High
32	12.89	No	High
33	14.11	Yes	High
34	14.84	No	High
35	36.81	Yes	High

Source: Fleming et al. (1980).

cardio-vascular diseases, such as the Beta-blocker Heart Attack Trial (BHAT, 1982). In clinical trials, however, it is more common to use the survival function for the disease with a high mortality rate. The survival function is the probability that a subject survives longer than y , namely

$$\begin{aligned}
 S(y) &= 1 - F(y) \\
 &= P(\text{a subject survives longer than } y) \\
 &= P(Y > y), \quad 0 < y < \infty. \tag{10.2.2}
 \end{aligned}$$

Hence $S(y)$ is a nonincreasing function of time y , and $S(0) = 1$ and $S(\infty) = 0$. As a result the probability density function (pdf) of Y is the probability of failure over a very small time interval,

$$\begin{aligned} f(y) &= \lim_{\Delta y \rightarrow 0} P(\text{a subject fails between } y \text{ and } y + \Delta y)/\Delta y \\ &= \lim_{\Delta y \rightarrow 0} P(y < Y \leq y + \Delta y)/\Delta y, \end{aligned} \quad (10.2.3)$$

Since the pdf of Y is the probability of failure over a time interval, it is greater than or equal to 0 if y is greater or equal to 0. It equals 0 if y is less than 0. In addition, the area under the curve $f(y)$ is equal to 1. Comparing the pdf given in (10.2.3) and the cdf given in (10.2.1) or survival function defined in (10.2.2) reveals that $f(y)$ is the derivative of $F(y)$, which equals minus of the derivative of the survival function $S(y)$. The hazard function, denoted as $h(y)$, is the instantaneous death rate, which is the conditional probability that a subject fails over the next instant given that the subject has survived up to the beginning of the interval. Mathematically it can be expressed as

$$\begin{aligned} h(y) &= \lim_{\Delta y \rightarrow 0} P(\text{a subject fails between } y \text{ and } y + \Delta y | \text{the subject survives to } y)/\Delta y \\ &= \lim_{\Delta y \rightarrow 0} P(y < Y \leq y + \Delta y | Y > y)/\Delta y \\ &= \frac{f(y)}{1 - F(y)} \\ &= \frac{f(y)}{S(y)}. \end{aligned} \quad (10.2.4)$$

Two hazard functions $h_1(y)$ and $h_2(y)$ are said to be proportional if

$$h_1(y) = \lambda h_2(y) \quad \text{for all } y > 0, \quad (10.2.5)$$

where λ is a constant, $\lambda > 0$. From (10.2.4) and (10.2.5), the relationship between the corresponding survival functions is given by

$$S_1(y) = [S_2(y)]^\lambda \quad \text{for all } y > 0. \quad (10.2.6)$$

As a result, if the ratio of two hazard functions is a constant, then the one survival function can be expressed as the other survival function to the λ th power.

If a predefined clinical event is observed in some subjects before the completion of the study, then their exact failure times are known. On the other hand, some subjects may withdraw prematurely without observing any occurrences of the event of interest due to some known or unknown reasons. Sometimes, the event does not occur for some subjects who completed the study. As a result, the time to the occurrence of the event is censored at the last known contact, and it is at least as long as the time from randomization to the time of the last contact. Let C denote the censoring time associated with the failure time Y . If C is greater than or equal to Y , then the survival time is actually observed. On the other hand, if the survival time is greater than the censoring time, then the survival time is not observed and is censored. As a result, the censored data for a subject consist of a pair of responses.

The first response is the observed time and the second is an indicator identifying whether the observed time is the survival time or was censored at the last contact. In other words, the data for the time to the occurrence of a predefined event obtained from n subjects of a clinical trial can be arranged as $(y_1, c_1), \dots, (y_n, c_n)$, where y_i is the observed time for subject i and

$$c_i = \begin{cases} 1 & \text{if } y_i \text{ is the survival time,} \\ 0 & \text{if } y_i \text{ is censored.} \end{cases} \quad (10.2.7)$$

This type of censoring mechanism is referred to as random censoring (Miller, 1981). Description for other types of censoring mechanism can be found in Miller (1981) and Lee and Wang (2003).

Since it is not easy to theoretically evaluate the true distribution of the survival time (Lee and Wang, 2003; Kalbfleisch and Prentice, 1980), in this chapter we will only cover Kaplan and Meier's nonparametric method for the survival function. For parametric methods, see Lee and Wang (2003) and Marubini and Valsecchi (1995). Let $y_{(1)} < \dots < y_{(K)}$ be the ordered distinct failure times when the event occurs and d_k be the number of events at time $y_{(k)}$ and m_k be the number of censored observations in the interval $(y_{(k)}, y_{(k+1)})$, $k = 1, \dots, K$. The risk set just prior to the time $y_{(k)}$ consists of the subjects who still survive and whose survival time is not censored before $y_{(k)}$. Thus, under the assumption of independent censoring, the number of subjects in the risk set just prior to the time $y_{(k)}$, denoted by n_k , is given by

$$n_k = (d_k + m_k) + \dots + (d_K + m_K), \quad k = 1, \dots, K.$$

The Kaplan-Meier nonparametric estimation of the survival function at time y is given by

$$\hat{S}(y) = \prod_{y_{(k)} < y} \left(1 - \frac{d_k}{n_k}\right). \quad (10.2.8)$$

The Kaplan-Meier estimate provides a straightforward yet intuitive interpretation of the survival function. The probability that a patient is alive (or without the event) at $y_{(k)}$ is equal to the conditional probability that the patient is alive at $y_{(k)}$, given that this patient survived through all proceeding time points when other patients failed times the probability that this patient survived all previous time points. In other words,

$$\begin{aligned} \hat{S}(y_{(k)}) &= P(\text{surviving at } y_{(k)}) \\ &= P(\text{surviving through } y_{(1)}, y_{(2)}, \dots, y_{(k-1)}, y_{(k)}) \\ &= P(\text{surviving } y_{(k)} | \text{surviving through } y_{(1)}, y_{(2)}, \dots, y_{(k-1)}) \\ &\quad \times P(\text{surviving through } y_{(1)}, y_{(2)}, \dots, y_{(k-1)}) \\ &= P(\text{surviving } y_{(k)} | \text{surviving through } y_{(1)}, y_{(2)}, \dots, y_{(k-1)}) \\ &\quad \times P(\text{surviving } y_{(k-1)} | \text{surviving through } y_{(1)}, y_{(2)}, \dots, y_{(k-2)}) \\ &\quad \times \dots \times P(\text{surviving } y_{(2)} | \text{surviving through } y_{(1)}) \\ &\quad \times P(\text{surviving } y_{(1)}). \end{aligned} \quad (10.2.9)$$

From (10.2.9), we see that the Kaplan-Meier estimates of survival are the same as those between two adjacent observed failure times. Then, the graph of the Kaplan-Meier estimates

Table 10.2.2 Data Layout for Computation of Kaplan-Meier Survival Function

Ordered Distinct Event Time	Number of Events	Number Censored in $[y_{(k)}, y_{(k+1)}]$	Number in Risk Set	$\hat{S}(y)$
$y_{(0)} = 0$	$d_0 = 0$	m_0	n_0	1
$y_{(1)}$	d_1	m_1	n_1	$1 - d_1/n_1$
$y_{(2)}$	d_2	m_2	n_2	$(1 - d_1/n_1)(1 - d_2/n_2)$
\vdots	\vdots	\vdots	\vdots	\vdots
$y_{(K)}$	d_K	m_K	n_K	$(1 - d_1/n_1)(1 - d_2/n_2) \dots (1 - d_K/n_K)$

must be a step function. According to the above discussion, the data layout for computing the Kaplan-Meier nonparametric estimates of survival consists of five columns as demonstrated in Table 10.2.2. The first column shows the ordered distinct failure times, the second column the number of events, the third column the number of censored observations, and the fourth column the number of subject at risk prior to $y_{(k)}$. The last column gives the Kaplan-Meier estimates of the survival function which are a product of the proportions of the number of surviving subjects at $y_{(k)}$ and of all proceeding distinct failure times.

The variance of the Kaplan-Meier estimate at time y can be found using Greenwood's formula:

$$v[\hat{S}(y)] = [\hat{S}(y)]^2 \sum_{y_{(k)} < y} \frac{d_k}{n_k(n_k - d_k)}. \quad (10.2.10)$$

It follows that a large-sample $(1 - \alpha)100\%$ confidence interval for $S(y)$ can be obtained as

$$\hat{S}(y) \pm Z(\alpha/2) \sqrt{v[\hat{S}(y)]}. \quad (10.2.11)$$

The median survival time can then be estimated as

$$q_{0.5} = \min\{y : 1 - \hat{S}(y) \geq 0.5\}. \quad (10.2.12)$$

A large-sample $(1 - \alpha)100\%$ confidence interval for the median survival time can be constructed similarly.

Example 10.2.1 The data of time to progression (in months) for 15 patients with limited stage II or stage IIIA low-grade ovarian cancer given in Table 10.2.1 are used to illustrate the computation of the Kaplan-Meier survival function in finding the probability of the patients who did not progress. In the data given in Table 10.2.3, ovarian cancer progressed in 6 out of 15 patients. At each of the first five progression time points, 0.92, 2.93, 5.76, 6.41, and 10.16 months, there was only one patient with occurrence of the event of cancer progression. The numbers of the patients in the risk set just prior to the progression time were 15, 14, 13, 12, and 11, respectively. However, between 10.16 and 15.20 months, there were 4 patients whose progression times were censored. At 15.20 months, the cancer progressed in one patient. As a result, the number of patients in the risk set just prior to 15.20 months became 6. Since the progression times for the last 5 patients were censored

Table 10.2.3 Computation of Kaplan-Meier Survival Function for Patients with Low-Grade Cancer in Table 10.2.1

Ordered Distinct Progression Time	Number of Events	Number of Censored in $[y_{(k)}, y_{(k+1)}]$	Number in Risk Set	$S(y)$
0	0	0	15	1
0.92	1	0	15	0.9333
2.93	1	0	14	0.8667
5.76	1	0	13	0.8000
6.41	1	0	12	0.7333
10.16	1	4	11	0.6667
15.20	1	1	6	0.5556

and were longer than 15.20 months, the probability of the patients who did not progress at or after 15.20 months can be estimated by (10.2.9) as

$$\hat{S}(15.20 \text{ or longer}) = \left(\frac{14}{15}\right) \left(\frac{13}{14}\right) \left(\frac{12}{13}\right) \left(\frac{11}{12}\right) \left(\frac{10}{11}\right) \left(\frac{5}{6}\right) = 0.5556.$$

The Kaplan-Meier estimates, the corresponding estimated large sample variance, and 95% confidence intervals are given in Table 10.2.4. The Kaplan-Meier estimates for patients with low-grade cancer are plotted in Figure 10.2.3 along with those for patients with well-differentiated cells. A large-sample variance of $\hat{S}(15.20 \text{ or longer})$ computed by the Greenwood's formula in (10.2.10) is 0.0206. It follows that the large-sample 95% confidence interval is (0.2744, 0.8367). However, since the probability of patients free of progression at or after 15.20 months is 0.5556 which is greater than 0.5, the median time of no progression cannot be estimated for the patients with low-grade cancer. On the other hand, for the patients with well-differentiated cancer, the shortest time for the estimated probability of patients free of progression greater than 50% is 12.14 months which is the estimated median time of no progression. The corresponding large-sample 95% confidence interval is from 9.57 to 12.57 months.

10.3 COMPARISON BETWEEN SURVIVAL FUNCTIONS

In clinical trials, the most important inference regarding censored data is to determine whether the experimental therapy can reduce the mortality rate or improve the survival rate compared to the placebo or standard treatment. Therefore, statistical hypotheses are typically expressed as

$$\begin{aligned} H_0: S_1(y) &= S_2(y), \\ \text{vs. } H_a: S_1(y) &\neq S_2(y). \end{aligned} \tag{10.3.1}$$

One commonly used method for comparing two survival functions is the logrank test. The logrank test is considered to be the most powerful test for the alternative hypothesis that hazard functions are proportional, namely,

$$\begin{aligned} H_0: S_1(y) &= S_2(y), \\ \text{vs. } H_a: S_1(y) &= [S_2(y)]^\lambda. \end{aligned} \tag{10.3.2}$$

Table 10.2.4 Kaplan-Meier Survival Rates for Patients with Ovarian Carcinoma by Low-Grade or Well-Differentiated Cancer

Time in Months	Censored	Estimated Proportion	Estimated Variance	Standard Error	Lower 95% Limit	Upper 95% Limit
<i>TMT = Low Grade</i>						
0.92	No	0.93333	0.004148	0.06440	0.80710	1.00000
2.93	No	0.86667	0.007703	0.08777	0.69464	1.00000
5.76	No	0.80000	0.010666	0.10328	0.59758	1.00000
6.41	No	0.73333	0.013037	0.11418	0.50955	0.95712
10.16	No	0.66667	0.014814	0.12171	0.42811	0.90523
12.40	Yes	0.66667				
12.93	Yes	0.66667				
13.85	Yes	0.66667				
14.70	Yes	0.66667				
15.20	No	0.55556				
23.32	Yes	0.55556				
24.47	Yes	0.55556				
25.33	Yes	0.55556				
36.38	Yes	0.55556				
39.67	Yes	0.55556				

Table 10.2.4 (Continued)

Time in Months	Censored	Estimated Proportion	Estimated Variance	Standard Error	Lower 95% Limit	Upper 95% Limit
<i>TMR = High Grade</i>						
1.12	No	0.95000	0.002375	0.04873	0.85448	1.00000
2.89	No	0.90000	0.004500	0.06708	0.76852	1.00000
4.51	No	0.85000	0.006375	0.07984	0.69351	1.00000
6.55	No	0.80000	0.008000	0.08944	0.62470	0.97530
9.21	No	0.75000	0.009375	0.09682	0.56023	0.93977
9.57	No	0.70000	0.010500	0.10247	0.49916	0.90084
9.84	Yes	0.70000				
9.87	Yes	0.70000				
10.16	No	0.64167	0.011942	0.10928	0.42748	0.85585
11.55	No	0.58333	0.012962	0.11385	0.36018	0.80648
11.78	No	0.52500	0.013562	0.11646	0.29675	0.75325
12.14	No	0.40833	0.013497	0.11618	0.18063	0.63604
12.17	No	0.35000	0.012833	0.11328	0.12797	0.57203
12.34	No	0.29167	0.011747	0.10838	0.07923	0.50410
12.57	No	0.23333	0.010240	0.10119	0.03499	0.43167
12.89	No	0.17500	0.007972	0.08929	0.00000	0.35370
14.11	Yes	0.17500	0.001993	0.04464	0.00000	0.23813
14.84	No	0.08750				
36.81	Yes	0.08750				

Source: Fleming et al. (1980).

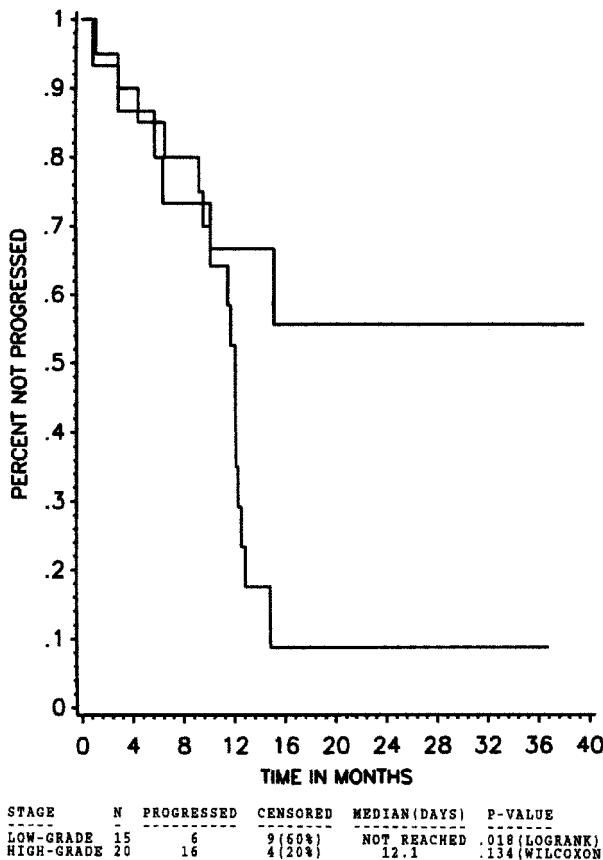


Figure 10.2.3 Distribution of Kaplan-Meier survival estimates on time to progression.

As was discussed in Section 9.5, the logrank test for comparing two independent samples is the same as the Mantel-Haenszel statistic for combining results from different strata. For logrank test statistic, a 2×2 contingency table by treatment and occurrence of events can be formed at each of the K distinct event times, $y_{(1)} < \dots < y_{(K)}$ (see Table 10.3.1). As was mentioned in Section 9.3, under the assumption of a hypergeometric distribution, the conditional expected number of patients with occurrence of events for d_{1k} at the event time $y_{(k)}$

Table 10.3.1 Data Structure of Comparing Two Survival Functions at $y_{(k)}$ by the Logrank Method

Treatment	Status		
	Event	No Event	Total
Test drug	d_{1k}	$n_{1k} - d_{1k}$	n_{1k}
Placebo	d_{2k}	$n_{2k} - d_{2k}$	n_{2k}
	d_k	$n_k - d_k$	n_k

Note: $k = 1, \dots, K$.

for the test drug is given by

$$e_{1k} = \frac{n_{1k}d_k}{n_k}, \quad (10.3.3)$$

and the conditional variance of d_{1k} is given as

$$v_{1k} = \frac{n_{1k}n_{2k}d_k(n_k - d_k)}{n_k^2(n_k - 1)}, \quad k = 1, \dots, K. \quad (10.3.4)$$

It turns out that the following logrank test statistic is the same as Mantel-Haenszel statistic given in (9.5.2):

$$\begin{aligned} X_{LR} &= \frac{\left[\sum_{k=1}^K (d_{1k} - e_{1k}) \right]^2}{\sum_{k=1}^K v_{1k}} \\ &= \frac{(d_1 - e_1)^2}{v_1}, \end{aligned} \quad (10.3.5)$$

where

$$\begin{aligned} d_1 &= \sum_{k=1}^K d_{1k}, \\ e_1 &= \sum_{k=1}^K e_{1k}, \\ v_1 &= \sum_{k=1}^K v_{1k}. \end{aligned}$$

Table 10.3.2 provides a summary for computation of logrank statistic. Under the null hypothesis of equal survival functions, the logrank statistic approximately follows a central chi-square distribution with one degree of freedom when sample size is moderate. Hence, the null hypothesis (10.3.1) is rejected at the α th significance level if

$$X_{LR} > \chi^2(\alpha, 1), \quad (10.3.6)$$

where $\chi^2(\alpha, 1)$ is the α th upper quantile of a central chi-square distribution with one degree of freedom. Peto et al. (1976) suggest that the relative hazard rate λ can be estimated as

$$\hat{\lambda} = \exp\left(\frac{d_1 - e_1}{v_1}\right), \quad (10.3.7)$$

Table 10.3.2 Computation of Logrank Statistic

Ordered Distinct Event Time	Observed Number of Events	Expected Number of Events	Difference	Variance
$y_{(1)}$	d_{11}	$e_{11} = n_{11}d_1/n_1$	$d_{11} - e_{11}$	v_{11}
$y_{(2)}$	d_{12}	$e_{12} = n_{12}d_2/n_2$	$d_{12} - e_{12}$	v_{12}
\vdots	\vdots	\vdots	\vdots	\vdots
$y_{(K)}$	d_{1K}	$e_{1K} = n_{1K}d_K/n_K$	$d_{1K} - e_{1K}$	v_{1K}
	d_1	e_1	$d_1 - e_1$	v_1

Note: $v_{1k} = n_{1k}n_{2k}d_k(n_k - d_k)/[n_k^2(n_k - 1)]$.

where “exp” denotes the natural exponentiation. Therefore, a large-sample $(1 - \alpha)100\%$ confidence interval for λ can be obtained as

$$\exp\{[d_1 - e_1]/v_1\} \pm Z(\alpha/2)/\sqrt{v_1}, \quad (10.3.8)$$

where $Z(\alpha/2)$ is the upper $(\alpha/2)$ th quantile of a standard normal distribution.

Sometimes, the randomization of patients is performed according to some predefined or natural stratified criteria such as gender or ST-segment elevation in assessment of recombinant hirubin with heparin for the treatment of acute coronary syndromes (GUSTO IIb, 1996). An extension of the logrank statistic is to combine results from strata. Suppose that there are a total of H independent strata. For each strata, d_{1h} , e_{1h} , and v_h can be calculated according to (10.3.5), $h = 1, \dots, H$, then the Mantel-Haenszel technique can be applied to combine the results

$$X_{\text{MH}} = \frac{\left[\sum_{h=1}^H (d_{1h} - e_{1h}) \right]^2}{\sum_{h=1}^H v_{1h}} \quad (10.3.9)$$

The null hypothesis of (10.3.1) is rejected if (10.3.6) is true.

The logrank statistic is in fact a special case of a general class of nonparametric tests for comparing two survival functions which can be expressed in the form of

$$T = \sum_{k=1}^K w_k (d_{1k} - e_{1k}), \quad (10.3.10)$$

where w_k is the weight assigned at the event time $y_{(k)}$. When the weight in (10.3.10) is 1, the square of T is the numerator of logrank test statistic in (10.3.5). On the other hand, if the weight is the proportion of patients in the risk set at event time $y_{(k)}$ to the total sample size, namely n_k/n , $k = 1, \dots, K$, the resulting test statistic is Gehan's generalization of the Wilcoxon two-sample rank sum test to the censored data (Gehan, 1965a, 1965b). Since Gehan's test gives greater weights to differences that occur at the beginning of a trial, it is more powerful in the detection of differences evidently early in time (Lee, Desu, and Gehan, 1975). On the other hand, the logrank test is fully efficient (i.e., the most powerful test) when relative hazards between the two survival functions is a constant. However, it is more sensitive to differences between the two survival functions that occur late in time. Tarone and Ware (1977) suggest using the square root of the number of patients in the risk set just prior to $y_{(k)}$ as the weights for the statistic given in (10.3.10). Another generalization of Wilcoxon rank sum test suggested by Prentice (1978) is to use the following weights:

$$S(y) = \prod_{y_{(k)} < y} \left[1 - \frac{d_k}{n_k + 1} \right]. \quad (10.3.11)$$

Note that the logrank statistic for comparing two survival functions can be extended to trials for assessment of more than two treatments. Suppose that there are a total of I different treatments including the placebo which is designated as the last treatment with index I . Similar to the two independent treatment groups, an $I \times 2$ contingency table by treatment and the occurrence of event can be formed at each of the K distinct event times so that $y_{(1)} < \dots < y_{(K)}$, as illustrated in Table 10.3.3. Conditional on the marginal totals, the expected number of patients with event for d_{ik} are given by

$$e_{ik} = \frac{n_{ik} d_k}{n_k}, \quad k = 1, \dots, K, i = 1, \dots, I. \quad (10.3.12)$$

Table 10.3.3 Data Structure for comparing I Survival Functions at $y_{(k)}$ by Logrank Method

Treatment	Status		
	Event	No Event	Total
Test drug 1	d_{1k}	$n_{1k} - d_{1k}$	n_{1k}
Test drug 2	d_{2k}	$n_{2k} - d_{2k}$	n_{2k}
\vdots	\vdots	\vdots	\vdots
Test drug i	d_{ik}	$n_{ik} - d_{ik}$	n_{ik}
\vdots	\vdots	\vdots	\vdots
Placebo	d_{Ik}	$n_{Ik} - d_{Ik}$	n_{Ik}
	d_k	$n_k - d_k$	n_k

Therefore the sum of difference between the observed and expected number of events over all K distinct event time points for treatment i is then given by

$$\sum_{k=1}^K (d_{ik} - e_{ik}) = d_i - e_i, \quad (10.3.13)$$

where

$$d_i = \sum_{k=1}^K d_{ik} \quad \text{and} \quad e_i = \sum_{k=1}^K e_{ik}, \quad i = 1, \dots, I.$$

The covariance between treatment i and i' for the sum of difference between the observed and expected number of events is given as

$$v_{ii'} = \sum_{k=1}^K \frac{(n_k n_{i'k} \delta_{ii'} - n_{ik} n_{i'k}) d_k (n_k - d_k)}{n_k^2 (n_k - 1)}, \quad 1 \leq i, i' \leq I, \quad (10.3.14)$$

where $\delta_{ii'} = 1$, if $i = i'$; $\delta_{ii'} = 0$, otherwise. Let $\mathbf{d} = (d_1, \dots, d_{I-1})$, $\mathbf{e} = (e_1, \dots, e_{I-1})$, and $\mathbf{V} = \{v_{ii'}, 1 \leq i, i' \leq I-1\}$. Then, the null hypothesis that $S_1 = \dots = S_I$ is rejected if test statistic

$$X = (\mathbf{d} - \mathbf{e})' \mathbf{V}^{-1} (\mathbf{d} - \mathbf{e}) > \chi^2(\alpha, I-1), \quad (10.3.15)$$

where $\chi^2(\alpha, I-1)$ is the α th upper quantile of a central chi-square distribution with $I-1$ degrees of freedom.

Example 10.3.1 Again we use the data of the time to the progression (in months) in patients with ovarian cancer given in Table 10.2.1 to illustrate the computation of logrank statistic for testing hypotheses of (10.3.1) and to estimate the relative hazard rate λ as given in (10.3.7). First, we need to form all possible 2×2 contingency tables by treatment and the occurrence of event at each of distinct event times in the combined samples. For this data set, there are a total of 20 distinct time points recorded as the cancer progressed. For each 2×2 contingency table, compute the marginal totals n_{1k} , n_{2k} , d_k , $n_k - d_k$, and n_k , $k = 1, \dots, 20$. From the marginal totals, the expected number of events and variance of d_{1k} can be computed according to (10.3.3) and (10.3.4). For example, at the first time point (0.92 month) with

the occurrence of progression, $n_{11} = 15$, $n_{21} = 20$, $d_1 = 1$, $n_1 = 35$, and $n_1 - d_1 = 34$. It follows that

$$e_{11} = \frac{(1)(15)}{35} = 0.42857,$$

$$v_{11} = \frac{(15)(20)(1)(34)}{(35)^2(34)} = 0.24490.$$

The difference between d_{11} and e_{11} is equal to $1 - 0.42857 = 0.57143$. The intermediate results are given in Table 10.3.4. At the bottom of Table 10.3.4 are given the sums of $d_{1k} - e_{1k}$, and v_{1k} over all distinct event time points. As a result, the logrank test statistic is given by

$$\begin{aligned} X_{\text{LR}} &= \frac{\sum_{k=1}^{20} (d_{1k} - e_{1k})^2}{\sum_{k=1}^{20} v_{1k}} \\ &= \frac{(d_1 - e_1)^2}{v_1} \\ &= \frac{(-5.33279)^2}{5.10898} \\ &= 5.5664. \end{aligned}$$

Table 10.3.4 Computation of Logrank Test Statistic in Table 10.2.1 for Patients with Stage II or IIIA Ovarian Carcinoma

Time in Months	d_{1k}	d_{2k}	d_k	n_{1k}	n_{2k}	n_k	e_{1k}	$d_{1k} - e_{1k}$	v_{1k}
0.92	1	0	1	15	20	35	0.42857	0.57143	0.24490
1.12	0	1	1	14	20	34	0.41176	-0.41176	0.24221
2.89	0	1	1	14	19	33	0.42424	-0.42424	0.24426
2.93	1	0	1	14	18	32	0.43750	0.56260	0.24609
4.51	0	1	1	13	18	31	0.41935	-0.41935	0.24350
5.76	1	0	1	13	17	30	0.43333	0.56667	0.24556
6.41	1	0	1	12	17	29	0.41379	0.58621	0.24257
6.55	0	1	1	11	17	28	0.39286	-0.39286	0.23852
9.21	0	1	1	11	16	27	0.40741	-0.40741	0.24143
9.57	0	1	1	11	15	26	0.42308	-0.42308	0.24408
10.16	1	1	2	11	12	23	0.95652	0.04348	0.47637
11.55	0	1	1	10	11	21	0.47619	-0.47619	0.24943
11.78	0	1	1	10	10	20	0.50000	-0.50000	0.25000
12.14	0	2	2	10	9	19	1.05263	-1.05263	0.47091
12.17	0	1	1	10	7	17	0.58824	-0.58824	0.24221
12.34	0	1	1	10	6	16	0.62500	-0.62500	0.23438
12.57	0	1	1	9	5	14	0.64286	-0.64286	0.22959
12.89	0	1	1	9	4	13	0.69231	-0.69231	0.21302
14.84	0	1	1	6	2	8	0.75000	-0.75000	0.18750
15.20	1	0	1	6	1	7	0.85714	0.14286	0.12245
Sum	6	16	22					-5.33279	5.10898

The corresponding p -value is given by 0.0183. Thus, we reject the null hypothesis that the survival function for the patients with low-grade ovarian cancer is the same as that with well-differentiated ovarian cancer at the 5% level of significance.

An estimate of the relative hazard ratio can be obtained as

$$\begin{aligned}\hat{\lambda} &= \exp\left[\frac{d_1 - e_1}{v_1}\right] \\ &= \exp\left[\frac{-5.33279}{5.10898}\right] \\ &= \exp(-1.04381) \\ &= 0.35231.\end{aligned}$$

The corresponding large-sample 95% confidence interval for the relative hazard λ can also be obtained as

$$\begin{aligned}\exp\{-1.04381 \pm 1.96/\sqrt{5.10898}\} &= \{\exp(-1.91093), \exp(-0.17668)\} \\ &= (0.14794, 0.83805).\end{aligned}$$

Note that both the point estimate and the corresponding large-sample interval estimate for the relative hazard rate λ are smaller than 1. Therefore, we can conclude that the number of occurrences of progression among patients with low-grade cancer is statistically lower than among those with well-differentiated cancer. In addition, the time to progression for patients with low-grade cancer is also statistically longer than that with well-differentiated cancer. This can further be verified that if the weights at each distinct progression time point are chosen to be n_k/n , $k = 1, \dots, 20$. The Gehan test statistic is -84 with a variance of 3146. It follows that the corresponding chi-square value is 2.2428 with a p -value of 0.1342. Therefore, according to the Gehan's generalized Wilcoxon rank test, we fail to reject the null hypothesis of equal survival functions. As shown in Figure 10.2.3, the difference in survival functions between the patients with low-grade and well-differentiated cancers becomes evident only after 12 months. Therefore, as we learned earlier, the Gehan test is not as powerful as the logrank test for the detection of this difference later in time.

10.4 COX'S PROPORTIONAL HAZARD MODEL

To provide a fair and unbiased assessment of efficacy and safety of a test drug based on censored data, the Cox's proportional hazard model has become a routine statistical method since it was introduced by Cox (1972). For example, the West of Scotland Coronary Prevention Study Group (1995) reported that the reduction in risk of nonfatal myocardial infarction or death from coronary heart disease with the cholesterol-lowering agent pravastatin at 40 mg per day is 31% (95% confidence interval: 17 to 43%, p -value < 0.0001) as compared to the placebo in men with moderate hypercholesterolemia and no history of myocardial infarction. The Cholesterol and Recurrent Events (CARE, 1996) trial investigators also reported that pravastatin at 40 mg per day provides a 24% reduction in risk of coronary events (95% confidence interval: 9 to 36%, p -value = 0.003) in patients with myocardial infarction but with an average cholesterol level. For both studies the reduction in risk and its corresponding 95%

confidence interval were estimated through the Cox's proportional hazard model. On the other hand, in a study for evaluation of methotrexate alone, sulfasalazine and hydroxychloroquine, or a combination of all three drugs in the treatment of rheumatoid arthritis (O'Dell et al., 1995), to obtain unbiased comparisons among three treatments, Cox's proportional hazard model was applied to adjust for differences in covariates at the study entry such as the erythrocyte sedimentation rate, the patient's global status, and the total joint score. Another example of application of the Cox's proportional hazard model is a prospective randomized trial comparing high-dose therapy and autologous bone marrow transplantation and conventional chemotherapy in patients with multiple myeloma (Attal et al., 1996). In addition to a comparison between treatments, Cox's proportional hazard model was used to identify the level of beta2-microglobulin in serum as one of the prognostic factors for the overall and event-free survivals.

Instead of a direct formulation of the survival function with a constraint between 0 and 1, the dependent response variable in the Cox's proportional hazard regression model is the hazard function at time y that can be expressed as the product of a baseline hazard function and a function of covariates. Let $\mathbf{x} = (x_1, \dots, x_p)'$ be a vector of p covariates collected from a subject. The general form of the proportional hazard regression model can be expressed as

$$h(y; \mathbf{x}) = h_0(y) \Omega(\mathbf{x}; \boldsymbol{\beta}), \quad (10.4.1)$$

where $h_0(y)$ is called the baseline hazard function and $\boldsymbol{\beta}$ is a vector of p unknown regression coefficients that relates to the hazard function at time y and p covariates.

Since the hazard and survival functions are all positive quantities, Cox (1972) suggests that function $\Omega(\mathbf{x}; \boldsymbol{\beta})$ be formulated in terms of exponentiation as follows:

$$\begin{aligned} h(y; \mathbf{x}) &= h_0(y) \exp\{\boldsymbol{\beta}' \mathbf{x}\} \\ &= h_0(y) \exp\left\{\sum_{i=1}^p \beta_i x_i\right\}. \end{aligned} \quad (10.4.2)$$

Note that the regression part in the proportional hazard function, $\exp\{\boldsymbol{\beta}' \mathbf{x}\}$, does not involve time t if all covariates are independent of time. The time-independent covariates are referred to as those collected before the initiation of the study such as demographic or baseline characteristics. The covariates collected during the study may vary at different visits, which are time-dependent variables. The baseline hazard function $h_0(y)$ is a function of time but not a function of the covariates. In addition, the functional form of the baseline hazard function $h_0(y)$ is unspecified. It does explicitly express the relationship between covariates and the hazard function in a parametric form. As a result, the Cox's proportional hazard regression model is a semiparametric statistical procedure.

Another important property of Cox's proportional hazard regression model is that the relative risk or hazard ratio, which is the ratio of the hazard at time y to the baseline hazard, is a function of covariates and does not vary with time if all covariates are time-independent, namely

$$\lambda(\mathbf{x}) = \frac{h(y; \mathbf{x})}{h_0(y)} = \exp\left\{\sum_{i=1}^p \beta_i x_i\right\}. \quad (10.4.3)$$

The proportionality of the Cox's proportional hazard model can also be expressed in terms of the survival function as

$$\begin{aligned} S(y) &= [S_0(y)]^{\lambda(x)} \\ &= [S_0(y)]^{\exp\left[\sum_{i=1}^p \beta_i x_i\right]}, \end{aligned} \quad (10.4.4)$$

where $S_0(y)$ is the survival function corresponding to the baseline hazard function $h_0(y)$.

For the moment suppose that the covariates of interest are the treatment indicator and gender. Let X_1 be the treatment indicator, which has a value of 1 if treatment is the test drug and has a value of 0 if treatment is the placebo. Similarly, X_2 represents gender, which has a value of 1 if a subject is a female and is 0 if a subject is a male. Then, model (10.4.2) can be written as

$$h(y; x_1, x_2) = h_0(y) \exp\{\beta_1 x_1 + \beta_2 x_2\}, \quad (10.4.5)$$

where β_1 measures treatment effect after adjustment of difference in gender and β_2 reflects the prognostic effect of gender, adjusted for the treatment effect. If we are interested in comparing the test drug and the placebo in female patients, the explicit expression of the hazard for a female subject assigned to receive the test drug at time t , according to (10.4.5), is given by

$$h(y; 1, 1) = h_0(y) \exp(\beta_1 + \beta_2).$$

Therefore, the relative risk with respect to the baseline hazard is

$$\lambda(1, 1) = \exp(\beta_1 + \beta_2).$$

The hazard corresponding to a female given the placebo is

$$h(y; 0, 1) = h_0(y) \exp(\beta_2)$$

with the respective relative risk

$$\lambda(0, 1) = \exp(\beta_2).$$

As a result, the relative risk of the test drug to the placebo for a female subject is then given by

$$\begin{aligned} \frac{\lambda(1, 1)}{\lambda(0, 1)} &= \frac{\exp(\beta_1 + \beta_2)}{\exp(\beta_2)} \\ &= \exp(\beta_1). \end{aligned} \quad (10.4.6)$$

We note that first, (10.4.6) only consists of the regression coefficient with respect to treatment assignment and is not contaminated with the covariate gender. As a result, it can be used to estimate the treatment effect, adjusted for difference in gender. Secondly, the relative risk of the test drug to the placebo for a female subject is a constant, which does not

vary with time. This is a key assumption of Cox's proportional hazard model, which implies that

$$S_{TF}(y) = [S_{PF}(y)]^{\exp(\beta_1)},$$

and

$$S_T(y) = [S_P(y)]^{\exp(\beta_1)}, \quad (10.4.7)$$

where $S_{TF}(y)$ is the survival function at time y for a female subject assigned to test drug and $S_{PF}(y)$ is the survival function at time y for a female subject given the placebo; $S_T(y)$ and $S_P(y)$ are similarly defined.

The second term of (10.4.7) implies that the relationship of survival functions between the test drug and the placebo for female subjects is the same as that of survival functions between the test drug and the placebo for any other covariates so long as the same constants are present in treatment groups. Let us denote $\exp(\beta_1)$ by λ . Since λ is a constant over time, the relationship of survival functions between the test drug and the placebo is the same as that specified in the alternative hypothesis of (10.2.6):

$$S_T(y) = [S_P(y)]^\lambda \quad \text{for all } y > 0, \quad (10.4.8)$$

where $\lambda = \exp(\beta_1)$. The relationship between $S_T(y)$ and $S_P(y)$ specified in (10.3.2) and (10.4.8) is referred to as Lehmann's alternative (Lehmann, 1953). This explains why the logrank test achieves full efficiency under the assumption that the hazard ratio (i.e., relative risk) between the test drug and the placebo is a constant.

The relative risk between the test drug and the placebo in (10.4.6) is expressed as the ratio of the relative risks between hazard of each group and the baseline hazard. It, however, can be written directly as the hazard ratio of the hazard function of the test drug to that of the placebo:

$$\begin{aligned} \frac{h(y; 1, 1)}{h(y; 0, 1)} &= \frac{h_0(y) \exp(\beta_1 + \beta_2)}{h_0(y) \exp(\beta_2)} \\ &= \exp(\beta_1). \end{aligned} \quad (10.4.9)$$

Note that the baseline hazard function of time appears in both numerator and denominator and therefore it is canceled out. As a result, the relative risk does not involve the baseline hazard function, and it is not a function of time if there are no time-dependent covariates. A statistical inference of relative risk and the regression coefficients, and of the corresponding survival functions, can still be made even though we do not specify the baseline hazard function. Cox's proportional hazard regression model is quite robust in the sense that it can adequately approximate the true but unknown parametric model. In addition, unlike the logistic regression for the occurrence of the events, Cox's proportional hazard model takes the survival time and censoring pattern into account. Hence, it is more efficient for the inference of the censored data than the logistic regression because Cox's proportional hazard model uses more information.

The likelihood function derived from Cox's proportional hazard model is based on the probabilities of the events of interest occurring in the subjects. It explicitly does not include the probabilities for subjects whose event times are censored. As a result the likelihood, since

it does not consist of probabilities for all subjects. Let $y_{(1)} < \dots < y_{(K)}$ be the ordered distinct times when the event occurs and $R(y_{(k)})$ be the risk set just prior to time $y_{(k)}$. Conditional on the risk set $R(y_{(k)})$, the probability of the occurrence of the event for subject j is given as

$$\frac{\exp\left(\sum_{i=1}^p \beta_i x_{ij}\right)}{\sum_{j \in R(y_{(k)})} \exp\left(\sum_{i=1}^p \beta_i x_{ij}\right)}. \quad (10.4.10)$$

For example, in the data on time to progression of cancer for the patients with limited stage II or IIIA ovarian carcinoma, the indicator covariate X takes the value 1 for low-grade cancer and 0 for well-differentiated cancer. The hazard functions at time y are $h_0(y)$ for the patients with well-differentiated cancer and $h_0(y)\exp(\beta)$ for the patients with low-grade cancer. Therefore, $\exp(\beta)$ is the relative risk of progression between lower-grade and well-differentiated cancers. The first progression occurs after 0.92 month (28 days) on a patient with low-grade cancer. The corresponding risk set just prior to 0.92 month consists of 15 patients with lower-grade cancer and 20 patients with well-differentiated cancer. As a result, the probability that progression did occur at this time in a patient with lower-grade cancer is given by

$$L_1 = \frac{\exp(\beta)}{[20 + 15 \exp(\beta)]}.$$

The second event of progression occurred at 1.12 months on a patient with well-differentiated cancer. The risk set just prior to 1.12 months includes 14 patients with low-grade cancer and 20 patients with well-differentiated cancer. Therefore, the conditional probability of progression at 1.12 month on a patient with well-differentiated cancer is

$$L_2 = \frac{1}{[20 + 14 \exp(\beta)]}.$$

We can repeat the same process until the last progression time. The partial likelihood corresponding to the Cox's proportional hazard model is the product of the conditional probabilities of the occurrence of events at each of the K distinct event times which is given by

$$\begin{aligned} L &= L_1 \times L_2 \times \dots \times L_K \\ &= \prod_{k=1}^K L_k \\ &= \prod_{k=1}^K \left\{ \frac{\exp\left(\sum_{i=1}^p \beta_i x_{ij}\right)}{\sum_{j \in R(y_{(k)})} \exp\left(\sum_{i=1}^p \beta_i x_{ij}\right)} \right\}. \end{aligned} \quad (10.4.11)$$

Since the partial likelihood is constructed based on the subjects with the occurrence of the event at their respective event times, it takes the event time into account. The risk set $R(y_{(k)})$ at time $y_{(k)}$ includes the subjects whose event times are censored after $y_{(k)}$. In other words, the partial likelihood derived from Cox's proportional hazard model also contains some information of censored observations. Once the partial likelihood is derived, then the maximum likelihood estimates (MLEs) of unknown regression coefficients in the Cox's proportional hazard model can be obtained through the standard iterative method by treating the partial likelihood as the ordinary likelihood. Let b_i and $se(b_i)$ represent the MLE of β_i and its associated large-sample standard error, respectively. Then, we can reject the null

hypothesis that $H_0: \beta_i = 0$ in favor of the alternative hypothesis that $H_a: \beta_i \neq 0$ if the Wald statistic

$$W_i = [b_i/se(b_i)]^2 > \chi^2(\alpha, 1), \quad i = 1, \dots, p, \quad (10.4.12)$$

where $\chi^2(\alpha, 1)$ is the α th upper quantile of a central chi-square distribution with 1 degree of freedom. A large-sample $(1 - \alpha)100\%$ confidence interval for β_i can be obtained as

$$b_i \pm Z(\alpha/2)se(b_i). \quad (10.4.13)$$

The confidence interval given in (10.4.13) is based on the Wald statistics where the standard error is calculated through the information matrix evaluated at the estimated b_i . The other equivalent test for large samples is the score test proposed by Rao (1973) where the information matrix is evaluated at values of β_i specified in the null hypothesis. If the model only includes the treatment indicator and all events occur at distinct event times (i.e., no tied event times), then the score test derived under Cox's proportional hazard model is the same as that under the logrank test for a comparison between two or more survival functions. This indicates that the log-rank test is a fully effective alternative measure of proportional hazard.

Let \mathbf{x}_k be a vector of p covariates from subject k , $\mathbf{x}_{k'}$ be a vector of p covariates associated with subject k' , and $\mathbf{b} = (b_1, \dots, b_p)'$ be a vector of the p MLEs of unknown regression coefficients. Then the relative risk or hazard ratio between patient k and k' can be estimated as

$$\hat{\lambda} = \exp[(\mathbf{x}_k - \mathbf{x}_{k'})'\mathbf{b}]. \quad (10.4.14)$$

A large-sample $(1 - \alpha)100\%$ confidence interval for λ is then given by

$$\exp\left\{(\mathbf{x}_k - \mathbf{x}_{k'})'\mathbf{b} \pm Z(\alpha/2)\sqrt{(\mathbf{x}_k - \mathbf{x}_{k'})'\nu(\mathbf{b})(\mathbf{x}_k - \mathbf{x}_{k'})}\right\}, \quad (10.4.15)$$

where $\nu(\mathbf{b})$ is the estimated large-sample covariance matrix of \mathbf{b} .

The validity of the statistical inference for the Cox's proportional hazard regression model depends on the assumption of proportional hazards among different values of covariates. Hence it is important to check this assumption before the application of the model. Several methods are available for detection of possible violation of proportional hazard assumption (e.g., see Kleinbaum, 1996). Two of these methods will be introduced in this chapter. The first method is the graphical approach. Recall that

$$S(y) = [S_0(y)]^{\exp(\sum \beta_i x_i)}.$$

It follows that the minus logarithm of $S(y)$ becomes

$$-\ln[S(y)] = -\left\{\ln[S_0(y)]\exp\left[\sum \beta_i x_i\right]\right\}.$$

Note that $-\ln[S(y)]$ is positive. As a result, the minus logarithm of $-\ln[S(y)]$ is a linear function of $\sum \beta_i x_i$ with intercept $-\ln\{-\ln[S_0(y)]\}$, namely

$$-\ln\{-\ln[S(y)]\} = -\ln\{-\ln[S_0(y)]\} - \sum_{i=1}^P \beta_i x_i. \quad (10.4.16)$$

Suppose that the only covariate in the model is the treatment indicator, which takes value 1 for the test drug and 0 for the placebo. Then, for a patient assigned to the test drug, we have

$$-\ln\{-\ln[S_T(y)]\} = -\ln\{-\ln[S_0(y)]\} - \beta.$$

On the other hand, for patients in the placebo group,

$$-\ln\{-\ln[S_P(y)]\} = -\ln\{-\ln[S_0(y)]\}.$$

It follows that the difference in $-\ln\{-\ln[S(y)]\}$ between the test drug and the placebo is a constant independent of time as demonstrated below:

$$-\ln\{-\ln[S_T(y)]\} - (\ln\{-\ln[S_0(y)]\}) = \beta. \quad (10.4.17)$$

Under the assumption of proportional hazard between the test drug and the placebo, it can be seen from (10.4.17) that $-\ln\{-\ln[S(y)]\}$ should be parallel to each other and the distance between them is a constant. Therefore this property can be applied to check the assumption of proportional hazard. For a particular covariate we can obtain the Kaplan-Meier estimates of the survival function at each level of the covariate. Then, the minus logarithms of the minus logarithm of the Kaplan-Meier estimates can be plotted. If the curves are approximately parallel, then the proportional hazard assumption is not violated. If these curves cross each other or diverge, then the assumption is not met.

The graphical approach is the simplest and easiest way to verify the proportional hazard assumption. However, it only allows us to examine one covariate at a time. The graphical method cannot assess the proportional hazard assumption simultaneously for a group of covariates when there is more than one covariate in the model. Kleinbaum (1996) recommends that a conservative approach be taken in applying the graphical method. Unless there is a strong evidence of nonparallelism among the curves of $-\ln\{-\ln[S(y)]\}$, the proportional hazard assumption should not be rejected. However, the graphical method is still a subjective approach because it does not provide a statistical test for assessing the proportional hazard assumption for a group of covariates.

Inclusion of time-dependent covariates in Cox's proportional hazard model can provide another means of a formal statistical assessment of the proportional hazard assumption. For simplicity, we first consider one covariate for the treatment indicator X , which takes the value 1 for the test drug and 0 for the placebo. The proportional hazard assumption is then evaluated, in addition to X , by including the interaction between treatment indicator and some function of time $d(y)$ in the model. The resulting model is given by

$$h(y; x) = h_0(y)\exp\{\beta x + \theta xd(y)\}, \quad (10.4.18)$$

where θ measures the interaction between treatment and time (i.e., the differences in treatment effects at different time points), and $d(y)$ can be any function of time. In practice, there are many choices for function $d(y)$. For simplicity, we restrict our selection to simple monotone function of time such as $d(y) = y$ or $d(y) = \ln(y)$, for $y > 0$. Similar to the ordinary regression analysis, y or $\ln(y)$ needs to be centered to avoid unnecessary difficulties that may arise during the iterative processes used in searching the MLEs. If the hazard ratios remain fairly constant within some time intervals and are markedly different from one time interval to another, then the $d(y)$ can be chosen as a step function, such as $d(y) = 0$ if $y < y_0$, and $d(y) = 1$ if $y \geq y_0$, for

the two time intervals. A nonzero θ usually indicates that the hazard ratios are different over time at different values of the covariate under investigation. A positive (negative) value of θ implies that the hazard ratios increase (decrease) as time increases. As a result, the hypothesis of interest here is whether the regression coefficient is different from 0, namely

$$\begin{aligned} H_0: \theta &= 0, \\ \text{vs. } H_a: \theta &\neq 0. \end{aligned} \quad (10.4.19)$$

Let us denote $LL(\beta, \theta)$ as the log-likelihood function under the interaction model (10.4.18), and let $LL(\beta)$ be the log-likelihood function under the reduced model

$$h(y; x) = h_0(y)\exp\{\beta x\}.$$

The null hypothesis (10.4.19) is rejected if

$$-2[LL(\beta) - LL(\beta, \theta)] > \chi^2(\alpha, 1), \quad (10.4.20)$$

where $\chi^2(\alpha, 1)$ is the α th upper quantile of a central chi-square distribution with one degree of freedom. This approach can be extended to investigate the proportional hazard assumption for a set of covariates by including interaction terms between each covariate and some function of time:

$$h(y; \mathbf{x}) = h_0(y)\exp \left\{ \sum_{i=1}^p \beta_i x_i + \sum_{i=1}^p \theta_i x_i d_i(y) \right\}. \quad (10.4.21)$$

The corresponding global hypothesis for the assessment of the proportional hazard simultaneously for a group of covariates is then given by

$$\begin{aligned} H_0: \theta_1 &= \dots = \theta_p = 0, \\ \text{vs. } H_a: \theta_i &\neq 0 \text{ for at least one } i, 1 \leq i \leq p. \end{aligned} \quad (10.4.22)$$

Let $LL(\beta_1, \dots, \beta_p; \theta_1, \dots, \theta_p)$ be the log-likelihood function under the full model given in (10.4.21) and $LL(\beta_1, \dots, \beta_p)$ be the log-likelihood function under the reduced model without the interaction terms between covariates and time given in (10.4.2). Then the null hypothesis of no time–covariate interaction is rejected if

$$-2[LL(\beta_1, \dots, \beta_p) - LL(\beta_1, \dots, \beta_p; \theta_1, \dots, \theta_p)] > \chi^2(\alpha, p), \quad (10.4.23)$$

where $\chi^2(\alpha, p)$ is the α th upper quantile of a central chi-square distribution with p degrees of freedom.

Suppose that a total of H covariates does not satisfy the proportional hazard assumption, the stratified Cox proportional hazard model can be applied to control these H covariates. Suppose that each of these H covariates has q_h levels $h = 1, \dots, H$. As a result, the Q strata can be formed from these H covariates, where $Q = q_1 \times \dots \times q_H$. Instead of assuming the same baseline hazard function for all covariates, the stratified Cox proportional hazard model assumes that the baseline hazard function is the same within a stratum but is different for other strata. Therefore the stratified Cox proportional hazard model is given by

$$h_q(y; \mathbf{x}) = h_{0q}(y)\exp \left\{ \sum_{i=1}^p \beta_i x_i \right\}, \quad q = 1, \dots, Q. \quad (10.4.24)$$

For each stratum the partial likelihood can be formulated in the same manner as (10.4.11), which is denoted L_q , $q = 1, \dots, Q$. Since each stratum is mutually exclusive and consist of different subjects, the partial likelihood for the entire sample is the product of these Q partial likelihoods:

$$L = \prod_{q=1}^Q L_q. \quad (10.4.25)$$

Then the MLEs for the regression coefficients in (10.4.24) can be obtained by the standard method, and the statistical inference with respect to the covariates in the model is made in the usual way. The effects of the covariates included in the stratified Cox proportional hazard model are estimated after adjustment for other covariates including those used to form the strata. However, the effects of the covariates used for stratification cannot be estimated.

The stratified Cox's proportional hazard model in (10.4.24) assumes that the regression coefficients are the same for all strata. In other words, there is no interaction between the p covariates included in the model and those H covariates used for stratification. The stratified Cox proportional hazard model with interaction terms is given by

$$h_q(y; x) = h_{0q}(y) \exp \left\{ \sum_{i=1}^p \beta_{iq} x_i \right\}, \quad q = 1, \dots, Q. \quad (10.4.26)$$

There are a total of pQ regression coefficients in model (10.4.26), while model (10.4.24) includes p regression coefficients. As a result, the global null hypothesis of no interaction between the covariates in the model and covariates forming strata can be performed by the usual twice of the minus difference in log-likelihood functions between (10.4.24) and (10.4.26) with $p(Q - 1)$ degrees of freedom.

Example 10.4.1 In this example, we will once again use the data of progression time for patients with limited stage II or stage IIIA ovarian cancer to illustrate the application of Cox's proportional hazard model. The treatment effect is estimated as 1.1186 with a standard error of 0.4771 if treatment indicator takes the value 1 for patients with well-differentiated cancer and 0 for patients with low-grade cancer. The corresponding Wald statistics is 5.0652 with a p -value of 0.0224. Note that the score test statistic is 5.512 with a p -value of 0.0189, which is very close to the 0.0183 obtained by the logrank test in Example 10.3.1. The reason for the small discrepancy between the p -values is that there are two tied progression times at 10.16 months. A large-sample 95% confidence interval for the treatment effect is then equal to $1.1186 \pm (1.96)(0.4771) = (0.1445, 2.0928)$. As a result the relative risk (hazard ratio) of progression for a patient with well-differentiated cancer compared to others with low-grade cancer is estimated as $\exp(1.1186) = 3.0606$ with the corresponding 95% confidence interval $(\exp(0.1445), \exp(2.0928)) = (1.1554, 9.1076)$. Therefore, the risk of progression in patients with well-differentiated cancer is about three times as high as that with low-grade cancer. However, from the 95% confidence interval, the variability of the estimated relative risk is quite large.

Figure 10.4.1 shows the $-\ln\{-\ln[S(y)]\}$ curves for patients with low-grade and with well-differentiated cancer. Since the two curves intersect each other many times before 12 months, the proportional hazard assumption may be seriously in doubt. It is therefore suggested that a formal statistical test be performed, such as given in (10.4.20), to verify the assumption. Note that the median progression time is about 12 months. As can be seen from the figure, the two curves of minus logarithm of minus logarithm of survival function

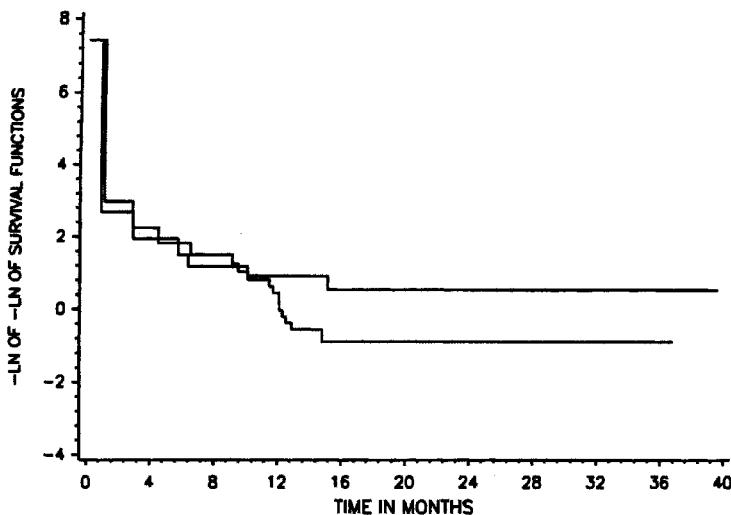


Figure 10.4.1 Doubled minus of logarithm of Kaplan-Meier survival estimates.

become roughly parallel to each other after 12 months. To reflect this phenomenon, we define the function as a function of time:

$$d(y) = \begin{cases} 0 & \text{for low-grade cancer,} \\ y - 12 & \text{well-differentiated cancer.} \end{cases}$$

The interaction θ between treatment and time in model (10.4.18) is estimated as 0.2438 with a 95% confidence interval from 0.0251 to 0.4625. Twice the minus of the log-likelihood is 123.528 for model (10.4.18) and is 128.610 for the reduced model. The difference is 5.082 with a p -value of 0.0248 which is significant at the 5% level. These results indicate that the risk for progression of ovarian cancer in a well-differentiated case of cancer is statistically greater than that of low-grade cancer, and this risk increases over time compared to the patients with low-grade cancer.

Example 10.4.2 Kalbfleisch and Street (1990) report the results from a clinical trial on cyclosporine and methotrexate (CSP + MTX) therapy versus methotrexate (MTX) alone followed by an infusion of marrow from an HLA-identical family member in patients with severe aplastic anemia. One of the primary endpoints is the time (in days) to severe (stage 2) acute graft versus host disease (AGVHD), death, or last contact. The data are reproduced in Table 10.4.1. The treatment indicator is coded as 0 for CSP + MTX and 1 for MTX. In addition to treatment, other covariates include age in years at the time of transplantation and laminar airflow isolation (LAF). Figure 10.4.2 gives the Kaplan-Meier survival estimates for the two treatments. The horizontal axis is truncated at 70 days because all events occurred before day 50 and event times of 41 patients (64%) were censored after day 49, more than 5 patients were censored with event times greater than 1,000 days. From Figure 10.4.2, it can be seen that the addition of CSP to MTX prolonged the time to severe AGVHD, or death over MTX alone. In addition Kalbfleisch and Street (1990) report that age is not only an important prognostic factor for the time to severe AGVHD or death but also interaction between treatment and age exists in favor of older patients (greater than or equal to 26 years older). On

Table 10.4.1 Time in Days to Severe (Stage 2) Acute Graft Versus Host Disease or Death for Patients with Severe Aplastic Anemia

Patient Number	Treatment	Age in Years	LAF	Time in Days	Censored
1	CSP + MTX	40	0	3	Censored
2	CSP + MTX	21	1	8	AGHVD
3	CSP + MTX	18	1	10	AGHVD
4	CSP + MTX	42	0	12	Censored
5	CSP + MTX	23	0	16	AGHVD
6	CSP + MTX	21	0	17	AGHVD
7	CSP + MTX	13	1	22	AGHVD
8	CSP + MTX	29	0	64	Censored
9	CSP + MTX	15	1	65	Censored
10	CSP + MTX	34	1	77	Censored
11	CSP + MTX	14	1	82	Censored
12	CSP + MTX	10	1	98	Censored
13	CSP + MTX	27	0	155	Censored
14	CSP + MTX	9	1	189	Censored
15	CSP + MTX	19	1	199	Censored
16	CSP + MTX	14	1	247	Censored
17	CSP + MTX	23	0	324	Censored
18	CSP + MTX	13	1	356	Censored
19	CSP + MTX	34	1	378	Censored
20	CSP + MTX	27	1	408	Censored
21	CSP + MTX	5	1	411	Censored
22	CSP + MTX	23	1	420	Censored
23	CSP + MTX	37	1	449	Censored
24	CSP + MTX	35	1	490	Censored
25	CSP + MTX	32	1	528	Censored
26	CSP + MTX	32	1	547	Censored
27	CSP + MTX	38	1	691	Censored
28	CSP + MTX	18	1	767	Censored
29	CSP + MTX	20	0	1111	Censored
30	CSP + MTX	12	0	1173	Censored
31	CSP + MTX	12	0	1213	Censored
32	CSP + MTX	29	0	1357	Censored
1	MTX	35	1	9	AGHVD
2	MTX	27	1	11	AGHVD
3	MTX	22	0	12	AGHVD
4	MTX	21	1	20	AGHVD
5	MTX	30	1	20	AGHVD
6	MTX	7	0	22	AGHVD
7	MTX	36	1	25	AGHVD
8	MTX	38	1	25	AGHVD
9	MTX	20	0	25	Censored
10	MTX	25	0	28	AGHVD
11	MTX	28	0	28	AGHVD
12	MTX	17	1	31	AGHVD
13	MTX	21	1	35	AGHVD
14	MTX	25	1	35	AGHVD
15	MTX	35	1	46	AGHVD

Table 10.4.1 (Continued)

Patient Number	Treatment	Age in Years	LAF	Time in Days	Censored
16	MTX	19	0	49	AGHVD
17	MTX	21	1	104	Censored
18	MTX	19	1	106	Censored
19	MTX	15	1	156	Censored
20	MTX	26	1	218	Censored
21	MTX	11	0	230	Censored
22	MTX	14	1	231	Censored
23	MTX	15	1	316	Censored
24	MTX	27	1	393	Censored
25	MTX	2	0	395	Censored
26	MTX	3	0	428	Censored
27	MTX	14	1	469	Censored
28	MTX	18	1	602	Censored
29	MTX	23	0	681	Censored
30	MTX	9	1	690	Censored
31	MTX	11	1	1112	Censored
32	MTX	11	0	1180	Censored

Note: CSP + MTX=cyclosporine and methotrexate, MTX = methotrexate, LAF = laminar airflow isolation.

Source: Kalbfleisch and Street (1990).

the other hand, LAF is not a significant prognostic factor, so it will be omitted from further discussion.

It seems that according to Figure 10.4.2, the hazard ratio of MTX to CSP + MTX increases over time. In order to investigate our conjecture of increasing hazard further and its relationship with age, as well as to follow the suggestion by Kalbfleisch and Street (1990), in addition to the original age we categorize age into three groups by the following two indicator variables AGE1 and AGE2:

$$AGE1 = \begin{cases} 1 & \text{if age is between 16 and 25 inclusively,} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$AGE2 = \begin{cases} 1 & \text{if age is greater than 25,} \\ 0 & \text{otherwise.} \end{cases}$$

The proportional hazard assumption is examined using the method described above by inclusion of time-dependent covariates in the model; these are the interaction between time-independent covariates and some function of time. For this example, we choose the function of time to be the identity function centered at the median of the following time of 173 days:

$$d(y) = y - 173.$$

Table 10.4.2 provides the twice minus of log-likelihood for the various models based on treatment, AGE1, AGE2, and their interactions with time using $d(y)$ as defined above. Comparison of the twice minus of log-likelihood from models with treatment only and

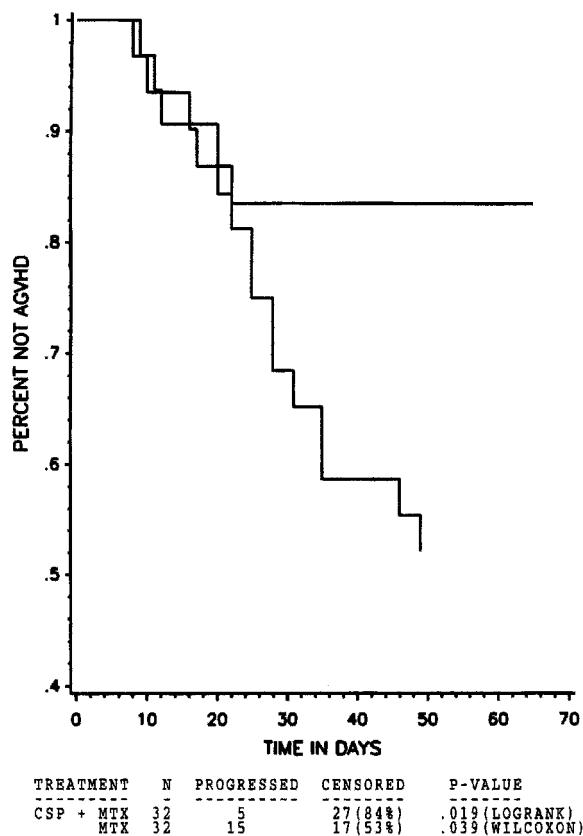


Figure 10.4.2 Distribution of Kaplan-Meier estimates on time to AGHVD.

with treatment and treatment-by-time interaction indicates an increase of 6.232 (152.587 – 146.355) with a *p*-value of 0.0119 in chi-square with one degree of freedom. Therefore, the hazard ratio between two treatments is not constant over time. In addition the regression coefficient corresponding to the interaction is estimated as 0.1582, which is positive. As a result the relative risk of progression for MTX to CSP + MTX increases over time. Both AGE1 and AGE2 are also significant prognostic factors because they increase the likelihood by an amount of 9.433 from the model of treatment alone (152.587 – 143.154). Their joint contribution is statistically significant at the 5% level of significance (*p*-value < 0.01). It is also of interest to check whether the interaction between treatment and time can be ignored after the inclusion of treatment, AGE1, and AGE2 in the model. The difference in the twice minus of the log-likelihood between the two models is 6.766 (143.154 – 136.388) with a *p*-value of 0.0093. To examine whether interactions between AGE1 and AGE2 with time exist, we test the joint contribution of two time-dependent covariates. The twice minus of the log-likelihood function is 135.744 which slightly increases by an amount of 0.644 with a *p*-value of 0.7247. As a result the model with treatment, AGE1, AGE2, and treatment-by-time interaction seems to be an adequate model. Table 10.4.3 provides the estimates, their corresponding large-sample standard errors for the model with treatment, AGE1, and AGE2 along with the model with inclusion of additional treatment-by-time interaction. Since the estimated coefficient for treatment-by-time interaction is

Table 10.4.2 Doubled Minus of the Log-likelihood for Various Models in Table 9.4.1 with Age Categorized into Three Groups

Model	<i>df</i>	-2LR	Chi-square
Without covariate		158.247	
TMT	1	152.587	5.661
TMT, TMT*g(y)	2	146.355	11.894
TMT, AGE1, AGE2	3	143.154	15.095
TMT, AGE1, AGE2, TMT*g(y)	4	136.388	21.857
TMT, AGE1, AGE2, AGE1*g(y)	4	143.108	15.140
TMT, AGE1, AGE2, AGE2*g(y)	4	143.148	15.099
TMT, AGE1, AGE2, TMT*g(y), AGE1*g(y)	5	136.303	21.946
TMT, AGE1, AGE2, TMT*g(y), AGE2*g(y)	5	135.905	22.234
TMT, AGE1, AGE2, AGE1*g(y), AGE2*g(y)	5	142.938	15.309
TMT, AGE1, AGE2, TMT*g(y), AGE1*g(y), AGE2*g(y)	6	135.744	22.503

0.1608, the increasing hazard ratio of MTX to CSP + MTX is statistically significant over time even after the adjustment for age. While the estimates for AGE1 and AGE2 for both models are quite similar between the two models, the estimate for a treatment effect is markedly different in magnitude. For the model without treatment-by-time interaction, the estimated treatment effect is 1.1651 with a large-sample standard error 0.5372. Hence the hazard ratio is estimated as 3.2062 with a large-sample 95% confidence interval from 1.1187 to 9.1890. On the other hand, for the model with inclusion of treatment-by-time interaction, the estimated treatment effect is inflated to 25.7644 with a large-sample standard error 12.7056. Hence the hazard ratio is estimated as 1.546×10^{11} with a large-sample 95% confidence interval from 2.3665 to 1.111×10^{22} . It seems that the addition of the product term between treatment and time causes instability in the estimation of the treatment effect.

For investigation of possible interaction between treatment and age, Table 10.4.4 presents the twice minus of the log-likelihood from six different models: successive significant contributions of age after treatment (chi-square = 5.169, *p*-value = 0.0229), treatment-by-time interaction after treatment and age (chi-square = 6.97, *p*-value = 0.0083), and treatment-by-age interaction after treatment, age, and treatment-by-time interaction (chi-square = 6.212, *p*-value = 0.0127). Because contributions of all four covariates, as assessed by log-likelihood, are all statistically significant at the 5% level, the resulting model is a four-parameter model. Table 10.4.5 displays the estimates of the regression coefficients and their large-sample

Table 10.4.3 Estimates of Models with Treatment, AGE1, AGE2, and Treatment-by-g(y) Interaction for Data in Table 10.4.1

Statistics	TMT	AGE1	AGE2	TMT*g(y)
Estimate	1.1651	1.9071	1.6776	
<i>se</i>	0.5372	0.7707	0.8097	
<i>p</i> -value	0.0301	0.0133	0.0383	
Estimate	25.7644	1.9331	1.7798	0.1608
<i>se</i>	12.7056	0.7710	0.8123	0.0813
<i>p</i> -value	0.0426	0.0122	0.0284	0.0480

Table 10.4.4 Doubled Minus of Log-likelihood for Various Models in Table 10.4.1 Without Age Categorized into Three Groups

Model	df	-2LR	Chi-square
Without covariate		158.247	
TMT	1	152.587	5.661
TMT, TMT*g(y)	2	146.355	11.894
TMT, AGE	2	147.418	10.929
TMT, AGE, TMT*g(y)	3	140.448	17.800
TMT, AGE, TMT*AGE	3	142.006	16.241
TMT, AGE, TMT*AGE, TMT*g(y)	4	134.236	24.012

standard errors for the four-parameter model and for the model without inclusion of treatment-by-time interaction, both obtained when age is both centered and not centered at the median age of 21 years old. Although the log-likelihood is the same whether age is centered or not, estimates of regression coefficients may be drastically different. For the model without treatment-by-time interaction, the estimated treatment effect is -1.6470 when age is not centered. On the other hand, when age is centered, the estimated treatment effect is 1.0678 , which is very close to the 1.1651 obtained from the model with treatment, AGE1, and AGE2. This example illustrates that the collinearity between age and treatment-by-age interaction with age not centered gives not only an erroneous estimate of the treatment effect but also a wrong sign. Other estimates, however, are not affected regardless of whether or not the age is centered. Furthermore the magnitude and sign of estimates for age and treatment-by-age interaction are very consistent for both three-parameter and four-parameter models. A negative estimate of age effect indicates that the risk of severe AGVHD or death is significantly smaller in older patients than younger patients. The hazard ratio of MTX to CSP + MTX increases over time as demonstrated by a significant positive estimated treatment-by-time interaction. Again the estimated treatment effect is inflated when treatment-by-time interaction is in the model. A significant treatment-by-time interaction indicates possible different baseline hazard functions between the two treatments. A stratified Cox proportional hazard model was fitted

Table 10.4.5 Comparison of Estimates of Regression Coefficients With and Without Age Centered at 21 Years Old in Table 10.4.1

Model	Statistics	TMT	AGE	TMT*AGE	TMT*g(y)
Age not centered	Estimate	-1.6470	-0.1641	0.1293	
	se	0.1293	0.1008	0.0569	
	p-value	0.2088	0.1035	0.0232	
Age centered	Estimate	1.0678	-0.1641	0.1293	
	se	0.5445	0.1008	0.0569	
	p-value	0.0499	0.1035	0.0232	
Age not centered	Estimate	24.1590	-0.1758	0.1398	0.1708
	se	12.6835	0.1022	0.0580	0.0812
	p-value	0.0568	0.0854	0.0159	0.0353
Age centered	Estimate	27.0956	-0.1758	0.1398	0.1708
	se	12.6770	0.1022	0.0580	0.0812
	p-value	0.0324	0.0854	0.0159	0.0353

**Table 10.4.6 Results of the Estimates of Age Effects
Stratified for Treatment in Table 10.4.1**

Statistics	AGE1	AGE2
Estimate	1.9309	1.7902
se	0.7712	0.8121
p-value	0.0123	0.0275

to the data using treatment as a stratification factor and AGE1 and AGE2 as covariates in the model. The results are given in Table 10.4.6. Their estimates and corresponding large-sample standard errors are very similar to those provided in Table 10.4.3 for both models without and with treatment-by-time interaction.

10.5 CALENDAR TIME AND INFORMATION TIME

At the planning stage the sample size is usually calculated based on information regarding the expected difference between treatments, its corresponding variability, and the desired statistical power for a predetermined risk of type I error. After the study is initiated, patients are enrolled to receive the treatment for a fixed length of time until he or she reaches either the end of the study or the time of analysis with survival as one of the primary endpoints. The Postmenopausal Estrogen/Progestin Interventions (PEPI), for example, is a trial funded by the U.S. National Institutes of Health to evaluate the effects of unopposed estrogen and combined estrogen-progestin therapy on four major cardiovascular disease risk factors in postmenopausal women (Espeland et al., 1995). These four risk factors include plasma HDL-cholesterol, systolic blood pressure, two-hour postoral glucose serum insulin, and plasma fibrinogen. Table 10.5.1 gives the five treatments to be assessed in the PEPI. In this trial 168 women per treatment for a total sample size of 840 women was determined to provide a minimum of 92% statistical power for detecting a difference of at least 5 mg/dl in mean change in HDL. The sample size was selected to control an overall type I error rate of 5% with Bonferroni adjustment for all pairwise comparisons among treatments. The sample size was also chosen to account for a 10% of lost to follow-up and 27% of dropouts. This study was conducted in seven clinical centers in the United States. Each center was expected to recruit and randomize 120 women. Based on the past experience, it was anticipated that enrollment and randomization of all women

Table 10.5.1 Treatments Evaluated in PEPI

Treatment	Estrogen	Progestin
1	Premarin: 0.625 mg daily	Placebo
2	Premarin: 0.625 mg daily	Provera: 10 mg days 1–12 Placebo: 13–28
3	Premarin: 0.625 mg daily	Provera: 25 mg daily
4	Premarin: 0.625 mg daily	Micronized progesterone: 200 mg days 1–12
5	Placebo	Placebo

Source: Espeland (1995).

could be completed in a 12-month recruiting period, beginning in December 1989. Although by the end of November 1990 recruiting activities were stopped, a decision was made to allow the women who entered the screening process to be enrolled. As a result the last woman was randomized in February 1991. It turned out that a total of 875 women were actually randomized and enrolled into the study. After randomization, each subject returned to one of the seven clinical centers after 3, 6, and 12 months for the first year and at 6-month intervals thereafter for a total of three years.

The length of actual trial of the PEPI was four years and three months. When there is a gap between the time of recruiting of a subject and the time of randomization after screening for eligibility, the number of subjects actually randomized is usually larger than the planned sample size, though for this study there was only about 4% more than the planned 840 subjects. In addition the length of the recruiting period for the PEPI was within the general limits of the expected time frames.

Since despite a well-prepared protocol and enrollment plan, lots of unexpected problems and logistic issues can occur during the recruiting period, the duration of the recruiting period of clinical trials is generally much longer than the anticipated length. The target of the planned sample size cannot be reached if appropriate actions are not taken. Therefore, after randomization of the first patient, the trial information starts to accumulate as time progresses. A full 100% of information should be accumulated by the time the last patient completes the scheduled follow-up.

As illustrated by the following examples, this is not always the case. In particular, for clinical trials with survival as one of the primary endpoints, the planned sample size is not even observed. For example, the Beta-blocker Heart Attack Trial (BHAT, 1982) was a randomized, double-blind, placebo-controlled trial sponsored by the National Heart, Lung, and Blood Institute, the U.S. National Institutes of Health to evaluate the effect of long-term use of propranolol on possible reduction of mortality of patients with a myocardial infarction 5 to 21 days within randomization. A total of 3837 patients were recruited between June 1978 and October 1980. The protocol was planned to have an average three-year follow-up period that was scheduled to end in June 1982. Since the primary endpoint was the mortality rate, the sample size was actually the expected deaths by June 1982, which was postulated to be 628 in the protocol. However, this number was never observed because the BHAT was terminated early in October 1981 after 318 deaths.

Clearly unexpected harmful events can force investigators to terminate the trial earlier. Another example was the Cardiac Arrhythmia Suppression Trial (CAST, 1989) which was stopped early because of a significantly higher number of deaths in patients treated with flecainide or encainide compared to the placebo group. The CAST was designed to enroll a total of 4,400 patients to achieve at least 80% power for detection of a 30% reduction in death from arrhythmia provided by the active drugs, under the assumption of a 11% mortality over three years for placebo. The preliminary report indicated that the CAST was not designed to prove that an antiarrhythmic drug can cause harm. However, the drugs did, and the trial was terminated early in April 1989. Both the original 425 arrhythmia events specified in the protocol and 300 lately revised in March 1989 for early termination were never observed because the CAST was stopped after 1455 patients received blind therapy and only fewer than 10% of the total expected events were observed.

Let N be the total sample size stated in the protocol for the provision of a desired power using a particular alternative hypothesis at a specified significance level. It is expected to recruit N subjects in the time interval $(0, T_c)$, where T_c is the maximum length of duration in calendar time for completion of the study. From the above discussion, T_c and N_c (the final

sample size of any clinical trials) are random variables. The relationship between the duration and sample size of a clinical trial can be explained by the concept of calendar time and information time (Lan and DeMets, 1989; Lan and Zucker, 1993). For simplicity this concept is illustrated through statistical inference for a population mean by the following hypotheses:

$$\begin{aligned} H_0: \mu &= \mu_0, \\ \text{vs. } H_a: \mu &\neq \mu_0. \end{aligned}$$

As was mentioned earlier, the variance of any estimates of μ obtained from the data is a measure of the precision of the estimate. The smaller variance is, the higher the precision in the estimation and the more information regarding parameter μ obtained. As a result the statistical information about μ provided by its estimator $\hat{\mu}$ computed from the data is defined as the inverse of the variance of $\hat{\mu}$. Let Y_1, \dots, Y_N be i.i.d. clinical responses with population mean μ and variance σ^2 . Then the sample mean \bar{Y} is the most commonly employed unbiased estimator for μ . Its variance is given by $\text{var}(\bar{Y}) = \sigma^2/N$. In other words, the information about the unknown population mean μ provided by the sample mean is defined as

$$I = \frac{N}{\sigma^2}, \quad (10.5.1)$$

which is the sample size times the inverse of the population variance. If $\sigma^2 = 1$, then the information about μ provided by the sample (Y_1, \dots, Y_N) is simply the sample size N .

The experimental units in most of clinical trials are subjects. As a result the information regarding the effectiveness and safety of the drug under study is also conceptually measured in terms of the number of patients who complete the study. The more subjects that complete the trial, the more clinical information is provided. The statistical information defined above therefore can also be interpreted as the clinical information. Thus the planned sample size specified in the protocol can be viewed as the minimum information required in order to achieve the desired power for detecting a minimum treatment effect at a predetermined risk of type I error.

As illustrated in the previous examples, almost all the clinical trials are longitudinal in nature in which responses of subjects are evaluated at prescheduled visits during the course of the study until either subjects complete the trial or they withdraw prematurely from the study for various reasons. The information defined above is based on the number of the patients who complete the study. In other words, the contribution to the information is 1 if the patient completes the study and is 0 otherwise. However, the information provided by the patients who discontinue the trial before the completion of the study is also valuable and useful for the assessment of efficacy, safety, and quality of life of the drug under study. The intention-to-treat analysis requires one to include all measurements at all time points by every patient because the information provided by the discontinued patients is not 0 but between 0 and 1.

Lan and Zucker (1993) give a definition of information for a particular subject for the inference of population mean slope under a linear random effects model as

$$I_i = \frac{1}{\sigma_\theta^2 \{1 + [R/SS(t_i)]\}}, \quad (10.5.2)$$

where

$$R = \frac{\sigma_e^2}{\sigma_\theta^2},$$

σ_e^2 is the variance for the within-subject error, σ_θ^2 is the variance for random subject-specific slope,

$$SS(t_i) = \sum_{j=1}^{J_i} (t_{ij} - \bar{t}_i)^2$$

and t_{ij} represent the time point j for subject i where the clinical endpoints were measured, $j = 1, \dots, J_i$; $i = 1, \dots, N$. The total information of the entire clinical trial I is then the sum of information from each individual as given by

$$I = \sum_{i=1}^N I_i.$$

$SS(t_i)$ in (10.5.2) is the correct sum of squares for the time points of subject i with the length of follow-up equal to J_i . In addition, $t_{i1} < \dots < t_{iJ_i}$. It follows that as J_i increases, more responses are measured and $SS(t_i)$ gets larger. Assuming that $\sigma_\theta^2 = 1$ if $SS(t_i) = 0$, then the individual information $I_i = 0$. However, as $SS(t_i)$ tends to infinity, the individual information becomes 1. The contribution by each subject to the total information depends on the number of measurements for the primary endpoints made by the individual. The number of measurements for the primary endpoints is, in general, proportional to the length of the duration that an individual stays in the trial. The longer a patient stays in the study, the more information is provided by the patient. The contribution by an individual patient is 1 if he or she completes the study.

As was indicated before, the entire duration of a clinical trial roughly consists of a fixed recruiting period plus either a fixed or an open-ended follow-up period. Subjects are enrolled into the study during different time points, which may be at a different entry rate during the recruiting period. This type of accrual of patients is referred to as the staggering entry. In addition patients may withdraw from the study prematurely during the follow-up period for some known or unknown reasons which may or may not be related to the treatments under the study. As a result, at a given calendar time point t , the information is different for each subject in the study. However, the total information can still be obtained at t by enumeration of the observed number of patients by the length of their durations in the follow-up period.

Suppose that a clinical trial plans to enroll a total of N patients with a maximum duration T_c in the calendar time scale. Let $n(t)$ be the number of subjects who complete the study at calendar time t . For the purpose of illustration, we use first definition of individual information which is 1 if a subject completes the trial. Therefore, the total information at calendar time t , $I(t)$ is just $n(t)$. $I(t)$ represents the amount of information accumulated by calendar time t . Similarly, the total information at the maximum duration T_c , $I(T_c)$, is $N(T_c)$ which is the maximum information expected to obtain at the maximum duration of the trial. It follows that the information time at calendar time t is defined as the proportion of the information available as of calendar t to the total information provided at the maximum duration T_c (Lan and Zucker, 1993). As a result, the relationship between calendar time and information time $s(t)$ is given by

$$s(t) = \frac{n(t)}{N(T_c)}, \quad 0 \leq t \leq T_c. \quad (10.5.4)$$

From (10.5.4) the information time is between 0 and 1. Increment of the number of subjects who complete the study is discrete. In other words, the information time $s(t)$ defined in (10.5.4) is also discrete despite the fact that calendar time is continuous. Let t_k be the

calendar time at which the k th subject completes the study, $k = 0, 1, \dots, N$. Then the corresponding information time is given as

$$\begin{aligned} s_k &= s(t_k) \\ &= \frac{n(t_k)}{N(T_c)}, \quad k = 0, \dots, N. \end{aligned} \quad (10.5.5)$$

We can transform the discrete information times defined in (10.5.5) into a continuous time by the following relationship:

$$s = \begin{cases} 0 & \text{if } s < s_1 = \frac{1}{N}, \\ s_k & \text{if } s \in [s_k, s_{k+1}). \end{cases} \quad (10.5.6)$$

In other words, the explicit expression for the relationship between calendar time and information time is given as

$$s(t) = \begin{cases} 0 & \text{if } t < t_1, \\ s_k & \text{if } t \in [t_k, t_{k+1}). \end{cases} \quad (10.5.7)$$

Formulation (10.5.7) states that the information time from calendar time t_k , when the k th subject completes the study, to just prior to t_{k+1} , when the $(k+1)$ th subject just about completes the study, is equal to $s_k = n(t_k)/N(T_c)$, the fraction of information available at calendar time t_k relative to the total information $N(T_c)$ at the maximum duration T_c .

The definition of information time and relationship between the information time and calendar time assumes that both the maximum duration T_c and maximum information $N(T_c)$ are known and that the total information at the maximum duration T_c is equal to the maximum information. However, as illustrated by the PEPI, CAST, and BHAT studies, both maximum information and maximum duration are random variables. If a clinical trial is allowed to continue until all N subjects, the predetermined sample size, have completed the scheduled follow-up period, then it is referred to as the maximum information trial. On the other hand, if a clinical trial is terminated at the maximum duration T_c , then it is referred to as the maximum duration trial. Since a maximum information trial continues until the last of all the predetermined number of patients completes the study, the total duration of the trial must be a random variable. Similarly for a maximum duration trial, the total information is a random variable.

Example 10.5.1 The results of BHAT (BHAT Investigators, 1982; Lan and DeMets, 1989) are used to illustrate the concept of calendar time and information time. The BHAT used the logrank test statistic given in (10.3.5) to compare the mortality of 1912 patients assigned to propranolol with that of 1921 patients given the placebo who were enrolled between June 1978 and October 1980. Unlike most test statistics for continuous or categorical responses, the information provided by the logrank test statistic is a function of the number of expected deaths. When the BHAT was designed, it was expected to have 628 deaths by June 1982, the scheduled end of the follow-up period. However, the maximum information of 628 deaths was never observed because the BHAT was terminated early in October 1981 due to convincing evidence of reduction in mortality provided by propranolol. The maximum information was revised to 400 when the data were available later in September 1982. The Data Monitoring Committee (DMC) met six times in May 1979, October 1979, March 1980, October 1980, April 1981, and October 1981. The corresponding

Table 10.5.2 Calendar Time and Information Time for Beta-Blocker Heart Attack Trial

Time of DMC	Calendar Time	Cumulative Proportion in Calendar Time	Cumulative Deaths	$D = 628$ Information Time	$D = 400$ Information Time
6/1978	0 month	0	0	0	0
5/1979	11 months	0.23	56	0.09	0.14
10/1979	16 months	0.33	77	0.12	0.19
3/1980	21 months	0.44	126	0.20	0.32
10/1980	28 months	0.58	177	0.28	0.44
4/1981	34 months	0.71	247	0.39	0.62
10/1981	40 months	0.83	318	0.51	0.80
6/1982	48 months	1.00	628	1.00	1.00

Source: Lan and DeMets (1989).

Note: DMC: Data-monitoring committee meeting.

number of deaths were 56, 77, 126, 177, 247, and 318. Since the BHAT enrolled the first patient in June 1978 and was originally scheduled to end the follow-up in June 1982, $T_c = 48$ months. $I(48)$ was originally planned to be 628 and later was revised to be 400. By the time when the BHAT was terminated by October 1981, the total duration of the trial in calendar time was 40 months which is about 83% of the maximum duration of 48 months specified in the protocol. There were a total of 318 deaths in October 1981. Hence $t_{318} = 40$ months. If the maximum information $I(48) = 628$ deaths, the corresponding information time at calendar time $t_{318} = 40$ months is $s(40) = 0.51$. On the other hand, if $I(48) = 400$, then $s(40) = 0.80$ which is quite close to 0.83, the cumulative proportion of 40 months with respect to the scheduled maximum duration of 48 months. Table 10.5.2 gives a summary of calendar times and information times for each of the six DMC meetings.

10.6 GROUP SEQUENTIAL METHODS

With almost no exception, clinical trials are longitudinal in nature, and it is not impossible to enroll and randomize all required patients on the same day either. The data of clinical trials are accumulated sequentially over time. It is therefore ethical and in the best interest of patients, as well as scientific and economic, to allow monitoring of information for management of the study and for decision making for possible early termination based on convincing evidence of either benefit or harm of the drug under investigation. The rationale for interim analyses of accumulating data in clinical trials was established by the Greenberg Report (Heart Special Project Committee, 1988) almost four decades ago. Since then, development of statistical methodology and decision processes for implementation of data monitoring and interim analyses for early termination has attracted a lot of attention. This section is to provide an overview of some of the most commonly employed statistical methods for interim analyses and data-monitoring process. Literature in this area is huge, and this overview is by no means comprehensive. Books on this field include Armitage (1975), Whitehead (1997), and Jennison and Turnbull (2000) for theoretical background and methodology development and Peace (1992) with emphasis on the biopharmaceutical applications. Armitage et al. (1969), Haybittle (1971), Peto et al. (1976), Pocock (1977), and O'Brien and Fleming (1979) provide some well-known group sequential methods. Lan

and DeMets (1983) introduce the alpha spending function. In addition, DeMets and Lan (1994) give a review of interim analyses through the alpha spending function. PMA Biostatistics and Medical Ad hoc Committee on interim analysis also published a position paper of the U.S. Pharmaceutical Manufacturing Association interim analysis in the pharmaceutical industry (PMA, 1993). Issues 5 and 6 of volume 12 of *Statistics in Medicine* (1993) published the proceedings of a workshop on *Practical Issues in Data Monitoring of Clinical Trials*, which was sponsored by the four institutes of the U.S. National Institutes of Health held at Bethesda, Maryland, on January 27–28, 1992. With respect to cancer trials, a workshop on *Early Stopping Rules in Cancer Clinical Trials* was held at Robinson College, Cambridge, England, on April 13–15, 1993. The proceedings for this workshop was published in Issue 13–14 of volume 13 of *Statistics in Medicine* (1994). In 2001, the U.S. FDA issued a draft guidance entitled *On the Establishment and Operation of Clinical Trials Data Monitoring Committee* (FDA, 2001c).

We consider a randomized, triple-blind, two parallel-group trial that compares an antihypertensive agent with a matching placebo. One of the primary endpoints is change from the baseline in diastolic blood pressure (mmHg). The hypotheses can be expressed as follows:

$$\begin{aligned} H_0: \mu_T &= \mu_P, \\ \text{vs. } H_a: \mu_T &\neq \mu_P, \end{aligned} \quad (10.6.1)$$

where μ_T and μ_P represent the population mean change from the baseline in diastolic blood pressure, respectively, for the test drug and the placebo. The intuitive test statistic for the null hypothesis of (10.6.1) is the Z-statistic which is given by

$$Z = \frac{\bar{Y}_T - \bar{Y}_P}{se(\bar{Y}_T) + se(\bar{Y}_P)}, \quad (10.6.2)$$

where \bar{Y}_T and \bar{Y}_P are the sample means in the change from the baseline in diastolic pressure computed from the patient receiving test drug and placebo, respectively, and $se(\bar{Y}_T)$ and $se(\bar{Y}_P)$ are the standard errors of \bar{Y}_T and \bar{Y}_P . When the sample size is at least moderate, the standard normal distribution provides an adequate approximation to the distribution of Z. Hence the null hypothesis of (10.6.1) is rejected at the 5% level of significance if the absolute value of the observed Z-value is greater than 1.96, which is the 2.5% upper quantile of a standard normal distribution.

The concept of the group sequential procedures is fairly simple. The number of planned interim analyses can be determined in advance and specified in the protocol. Let N be the total planned sample size with equal allocation to the two treatments. Suppose that K interim analyses are planned with equal increments of accumulating data. Then, we can divide the duration of the clinical trial into K intervals. Within each stage, the data of $n = N/K$ patients are accumulated. At the end of each interval, an interim analysis will be performed using Z-statistic, denoted by Z_i , in (10.6.2) with the data accumulated up to that point. Two decisions will be made based on the result of each interim analysis. First, the trial is continued if

$$|Z_i| \leq z_i, \quad i = 1, \dots, K-1, \quad (10.6.3)$$

where z_i are critical values known as the group sequential boundaries. However, we declare that we fail to reject the null hypothesis of (10.6.1) if

$$|Z_i| \leq z_i, \quad \text{for all } i = 1, \dots, K. \quad (10.6.4)$$

Table 10.6.1 Data Structure of Group Sequential Methods

Number of Interim Analysis (K)	Randomization		Information Time	Z-Statistic
	Test Drug	Placebo		
1	$n/2$	$n/2$	$1/K(n)$	Z_1
2	$2n/2$	$2n/2$	$2/K(2n)$	Z_2
3	$3n/2$	$3n/2$	$3/K(3n)$	Z_3
\vdots	\vdots	\vdots	\vdots	\vdots
K	$Kn/2$	$Kn/2$	$1(Kn)$	Z_K

Nevertheless, the null hypothesis is rejected, and we can terminate the trial if at any of the K interim analyses

$$|Z_i| > z_i, \quad i = 1, \dots, K. \quad (10.6.5)$$

For example, at the end of the first interval, an interim analysis is carried out with the data of n patients. If (10.6.3) is true, we continue the trial to the second planned interim analysis. Otherwise, we reject the null hypothesis, and we can stop the trial. After the trial continues to the end of the third interval, the accumulated data of $3n$ patients will be used for the third interim analysis. Data of all patients will be utilized for the final interim analysis after equation (10.6.3) is satisfied at all previous $K - 1$ interim analyses. The trial will be terminated at the final analysis regardless if (10.6.3) is true or not. If (10.6.4) is true at the final analysis, then we can declare that the data from the trial do not provide sufficient evidence to doubt the validity of the null hypothesis. Otherwise, the null hypothesis is rejected, and we can conclude that there is statistically significant difference in change from the baseline of diastolic blood pressure between the test drug and the placebo. The data structure for the group sequential procedures is illustrated in Table 10.6.1. As can be seen from Table 10.6.1, the interim analysis is scheduled after equal information of n patients is accumulated. Therefore the time scale for the interim analyses is the information time. For this example each interim analysis is planned based on information time at $1/K, 2/K, 3/K, \dots, (K - 1)/K$, and 1, which is called the K -stage group sequential procedure.

In contrast to the fixed sample where only one final analysis is performed, K analyses are carried out for the K -stage group sequential procedure. Suppose that the nominal significance level for each of the K interim analyses is still 5%. Then because of repeated testing based on the accumulated data, the overall significance level is inflated. In other words, when there is no difference in the change from the baseline of diastolic blood pressure between the test drug and the placebo, the probability of declaring at least one significance result increases due to K interim analyses. For the five planned interim analyses as shown in Table 10.6.2 (Armitage, et al., 1969), the overall type I error rate inflates to 14% instead of 5%. In other words, even though there is no difference between the test drug and the placebo, the odds of at least one false positive finding increase from 1 out of 20 to 1 in 7.

Various methods have been proposed to maintain the overall significance level at the pre-specified nominal level. One of the early methods was proposed by Haybittle (1971) and Peto et al. (1976). They proposed using 3.0 as the group sequential boundaries for all interim analyses except for the final analysis for which they suggested 1.96. In other words,

$$z_i = \begin{cases} 3.0 & \text{if } i = 1, \dots, K - 1, \\ 1.96 & \text{if } i = K. \end{cases}$$

**Table 10.6.2 Repeated Significance Tests
on Accumulative Data**

Number of Repeated Tests at 5% Level	Overall Significance Level
1	0.05
2	0.08
3	0.11
4	0.13
5	0.14
10	0.19
20	0.25
50	0.32
100	0.37
1000	0.53
∞	1.00

Source: Armitage et al. (1969).

Therefore their method can be summarized as follows:

- Step 1. At each of the K interim analyses, compute $Z_i, i = 1, \dots, K - 1$.
- Step 2. If the absolute value of Z_i crosses 3.0, then reject the null hypothesis and recommend a possible early termination of the trial; otherwise, continue the trial to the next planned interim analysis and repeat steps 1 and 2.
- Step 3. For the final analysis, use 1.96 for the boundary. The trial stops here regardless if the null hypothesis is rejected.

The Haybittle and Peto's method is very simple. However, it is a procedure with ad hoc boundaries that are independent of any planned interim analyses and stages. Pocock (1977) proposes different group sequential boundaries that depend on the number of planned interim analyses. However, his boundaries are constant at each stage of an interim analysis. For example, when the number of planned interim analyses is 5, Pocock (1977) suggests using 2.413 as the group sequential boundary for each of the five interims. However, if $K = 4$, then Pocock's boundary decreases to 2.361. The implementation of Pocock's group sequential procedure is the same as that of the Haybittle and Peto steps 1 to 3 above. The Pocock's group sequential boundaries are given in Table 10.6.3. Since limited information is included in the early stages of the interim analyses, O'Brien and Fleming (1979) suggest setting conservative boundaries for the interim analyses scheduled to be carried out in an early phase of the trial. Their group sequential boundaries not only depend on the number of the planned interim analysis but also are a function of their stages. As a result, the O'Brien-Fleming boundaries can be calculated as:

$$z_{iK} = \frac{c_K \sqrt{K}}{\sqrt{i}}, \quad 1 \leq i \leq K, \quad (10.6.6)$$

where c_K is the critical value for a total of K planned interim analyses also provided in Table 10.6.3. Again let us suppose that five planned interim analyses are scheduled. Then

Table 10.6.3 Group Sequential Boundaries

Number of Interim Analyses (K)	Group Sequential Boundaries		
	Pocock	O'Brien-Fleming	
		Value	Multiplier
1	1.960	$\sqrt{1/i}$	1.960
2	2.178	$\sqrt{2/i}$	1.978
3	2.289	$\sqrt{3/i}$	2.004
4	2.361	$\sqrt{4/i}$	2.024
5	2.413	$\sqrt{5/i}$	2.040
6	2.453	$\sqrt{6/i}$	2.053
7	2.485	$\sqrt{7/i}$	2.063
8	2.512	$\sqrt{8/i}$	2.072
9	2.535	$\sqrt{9/i}$	2.080
10	2.555	$\sqrt{10/i}$	2.086

Note: Two-sided: 0.05; One-sided: 0.025. $i = 1, \dots, K$.

Source: Modified from Jennison and Turnbull (1989, 1991).

$c_5 = 2.04$ and boundaries for each stage of these five interim analyses are given as

$$z_{i5} = \frac{2.04\sqrt{5}}{\sqrt{i}}, \quad 1 \leq i \leq 5.$$

For example, the O'Brien-Fleming boundary for the first interim analysis is equal to $(2.04)(\sqrt{5}) = 4.561$. The O'Brien-Fleming boundaries for the other four interim analyses can be similarly computed as 3.225, 2.633, 2.280, and 2.040, respectively. A graphical comparison of the group sequential boundaries among these three methods is given in Figure 10.6.1. It is evident from Figure 10.6.1 that the O'Brien-Fleming boundaries are so conservative, that the early trial results must be extreme for any prudent and justified decision-making in recommendation of a possible early termination when very limited information is available. On the other hand, in the late phase of the trial when the accumulated information approaches the required maximum information, their boundaries also become quite close to the critical value when no interim analysis has been planned. As a result, the O'Brien-Fleming method does not require a significant increase in the sample size for what has been already planned. Therefore the O'Brien-Fleming group sequential boundaries have become one of the most employed procedures for the planned interim analyses of clinical trials. Jennison and Turnbull (1989) suggest performing interim analyses by application of repeated confidence intervals. The repeated confidence intervals take the usual form of a confidence interval but with critical values replaced by the group sequential boundaries given in Table 10.6.3. The i th repeated confidence interval for $\mu_T - \mu_P$, the mean difference in change from the baseline of diastolic blood pressure between the test drug and the placebo, is given by

$$(\bar{Y}_T - \bar{Y}_P) \pm z_{iK}[se(\bar{Y}_T) + se(\bar{Y}_P)], \quad i \leq i \leq K, \quad (10.6.7)$$

where z_{iK} are the group sequential boundaries. The implementation of the repeated confidence interval approach is similar to the group sequential procedure described above. For a two-sided alternative, an early termination of the i th interim analysis is recommended if the i th repeated confidence interval does not contain 0. Otherwise, the trial continues. The repeated confidence interval approach is appealing because it combines aspects of sequential estimation and testing for a full exploration of the data at each interim analysis. The repeated

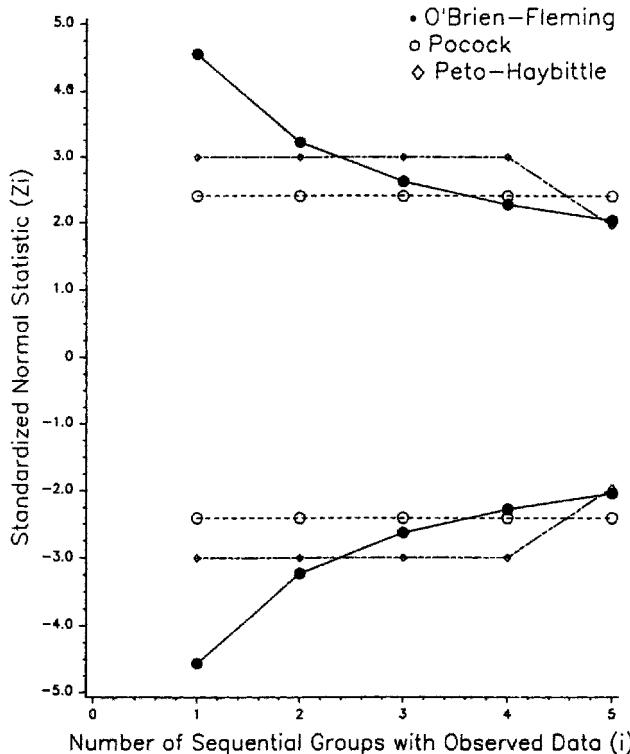


Figure 10.6.1 Two-sided 5% boundaries for Pocock, O'Brien-Fleming, and Peto-Haybittle methods. (Source: DeMets and Lan, 1994.)

confidence interval approach can also be applied to the equivalence problem (Jennison and Turnbull, 1993). Gould (1995b) extends the group sequential procedure to the bioequivalence problem.

The implementation of any group sequential procedures requires someone to specify the number of planned interim analyses in advance. It also dictates equal increments of information (number of patients) for each stage. As illustrated in Table 10.6.1, the time scale for scheduling interim analyses is the information time. However, the data-monitoring committee meetings for the review of the data and interim results are scheduled on calendar time. Unless the rate of accumulating information is uniform, it is quite difficult to convert information time to the calendar time for prescheduling the DMC meetings. DeMets and Lan (1994) describe in details this problem from their experiences with BHAT. Moreover it is not uncommon that an occurrence of some unexpected beneficial or harmful effect demands, for ethical and scientific reasons, changes in the frequency of the planned interim analyses and prescheduled DMC meetings. Under these circumstances the group sequential procedures become rather inflexible to limit the decisions and actions that the DMC must take for the best interest of the patients. Lan and DeMets (1983) in their landmark paper introduce the concept of the alpha spending function to overcome the drawbacks of traditional group sequential procedures.

In the standard group sequential procedures, the boundaries are determined by critical values so that the sum of probabilities of exceeding these values at the prescheduled discrete information time points is equal to the predetermined total probability of type I error. The idea of the alpha spending function proposed by Lan and DeMets (1983) is to spend

(i.e., distribute) the total probability of false positive risk as a continuous function of the information time. As a result, if the total information scheduled to accumulate over the maximum duration T is known, the boundaries can also be computed as a continuous function of the information time. This continuous function of the information time is referred to as the alpha spending function, denoted by $\alpha(s)$. The alpha spending function is an increasing function of information time. It is 0 when information time is 0, and it is equal to the overall significance level when information time is 1. In other words, $\alpha(0) = 0$ and $\alpha(1) = \alpha$. Let s_1 and s_2 be two information times, $0 < s_1 < s_2 \leq 1$. Also denote $\alpha(s_1)$ and $\alpha(s_2)$ as their corresponding value of alpha spending function at s_1 and s_2 . Then $0 < \alpha(s_1) < \alpha(s_2) \leq \alpha$. $\alpha(s_1)$ is the probability of type I error one wishes to spend at information time s_1 . With respect to $\alpha(s_1)$, the boundary $z(s_1)$ can then be computed by the following probability statement:

$$P\{Z(s_1) > z(s_1)\} = \alpha(s_1).$$

Suppose that the trial fails to terminate at s_1 and continues to accumulate information up to s_2 when we perform the second interim analysis. The cumulative false positive probability is given by

$$\begin{aligned}\alpha(s_2) &= P\{Z(s_1) > z(s_1) \text{ or } Z(s_2) > z(s_2)\} \\ &= P\{Z(s_1) > z(s_1)\} + P\{Z(s_1) \leq z(s_1) \text{ and } Z(s_2) > z(s_2)\} \\ &= \alpha(s_1) + P\{Z(s_1) \leq z(s_1) \text{ and } Z(s_2) > z(s_2)\}. \end{aligned}\quad (10.6.8)$$

Then $\alpha(s_2) - \alpha(s_1)$ is the proportion of the probability of type I error one is willing to allocate to the additional accumulation of information $s_2 - s_1$. The boundary $z(s_2)$ can be obtained through numerical integration by the following formulation:

$$\alpha(s_2) - \alpha(s_1) = P\{Z(s_1) \leq z(s_1) \text{ and } Z(s_2) > z(s_2)\}. \quad (10.6.9)$$

This relationship is given in Figure 10.6.2, which provides a graphical representation of the alpha spending function. Table 10.6.4 provides different forms of alpha spending functions whose boundaries for 5 interim looks are given in Table 10.6.5.

As an example, consider the O'Brien-Fleming group sequential procedure with a total of five planned interim analyses with equal increment of information for one-sided alternative

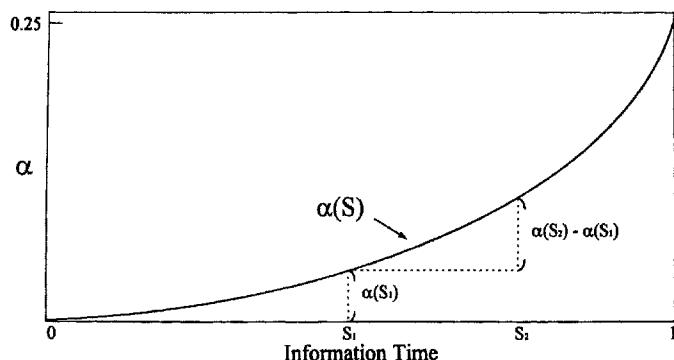


Figure 10.6.2 The alpha spending function $\alpha(s)$.

Table 10.6.4 Different Forms of Alpha Spending Functions

Approximation	
$\alpha_1(s) = 2\{1 - \Phi[z(\alpha/2)/\sqrt{s}]\}$	O'Brien-Fleming
$\alpha_2(s) = \alpha \ln[1 + (e - 1)s]$	Pocock
$\alpha_3(s) = \alpha s^\theta, \theta > 0$	Lan-DeMets-Kim
$\alpha_4(s) = \alpha[(1 - e^{-\zeta s})/(1 - e^{-\zeta})], \zeta \neq 0,$	Hwang-Shih

at the 2.5% overall significance level. The five boundaries are given as $z(0.2) = 4.56$, $z(0.4) = 3.23$, $z(0.6) = 2.63$, $z(0.8) = 2.28$, and $z(1.0) = 2.04$. The cumulative probabilities of type I error is given in Table 10.6.6. As a result, the cumulative false positive error rate at the five information time points are given as

$$\begin{aligned}
 P\{Z(0.2) > 4.56\} &= 2.6 \times 10^{-6} \\
 P\{Z(0.2) > 4.56 \text{ or } Z(0.4) > 3.23\} &= P\{Z(0.2) > 4.56\} + P\{Z(0.2) \leq 4.56 \text{ and } Z(0.4) > 3.23\} \\
 &= 2.6 \times 10^{-6} + 0.000574 = 0.0006, \\
 P\{Z(0.2) > 4.56 \text{ or } Z(0.4) > 3.23 \text{ or } Z(0.6) > 2.63\} &= P\{Z(0.2) > 4.56 \text{ or } Z(0.4) > 3.23\} \\
 &\quad + P\{Z(0.2) \leq 4.56 \text{ and } Z(0.4) \leq 3.23 \text{ and } Z(0.6) > 2.63\} \\
 &= 0.0006 + 0.0039 = 0.0045, \\
 P\{Z(0.2) > 4.56 \text{ or } Z(0.4) > 3.23 \text{ or } Z(0.6) > 2.63 \text{ or } Z(0.8) > 2.28\} &= P\{Z(0.2) > 4.56 \text{ or } Z(0.4) > 3.23 \text{ or } Z(0.6) > 2.63\} \\
 &\quad + P\{Z(0.2) \leq 4.56 \text{ and } Z(0.4) \leq 3.23 \text{ and } Z(0.6) \leq 2.63 \\
 &\quad \text{and } Z(0.8) > 2.28\} \\
 &= 0.0045 + 0.0080 = 0.0125, \\
 P\{Z(0.2) > 4.56 \text{ or } Z(0.4) > 3.23 \text{ or } Z(0.6) > 2.63 \text{ or } Z(0.8) > 2.28 &\text{ or } Z(1.0) > 2.04\} \\
 &= P\{Z(0.2) > 4.56 \text{ or } Z(0.4) > 3.23 \text{ or } Z(0.6) > 2.63 \\
 &\quad \text{or } Z(0.8) > 2.28\} \\
 &\quad + P\{Z(0.2) \leq 4.56 \text{ and } Z(0.4) \leq 3.23 \text{ and } Z(0.6) \leq 2.63 \\
 &\quad \text{and } Z(0.8) \leq 2.28 \text{ and } Z(1.0) > 2.04\} \\
 &= 0.0125 + 0.0125 = 0.0250
 \end{aligned}$$

Table 10.6.5 Examples of Boundaries by Alpha Spending Function

Interim Analysis(s)	O'Brien-Fleming	$\alpha_1(s)$	Pocock	$\alpha_2(s)$	$\alpha_3(s)[\theta = 1]$
1(0.2)	4.56	4.90	2.41	2.44	2.58
2(0.4)	3.23	3.35	2.41	2.43	2.49
3(0.6)	2.63	2.68	2.41	2.41	2.41
4(0.8)	2.28	2.29	2.41	2.40	2.34
5(1.0)	2.04	2.03	2.41	2.39	2.28

Note: Number of interim analyses = 5. Two-sided: 0.05; One-sided: 0.025.

Table 10.6.6 Cumulative Probability of Type I Error

Interim Analysis(s)	Group Sequential Boundaries					
	Pocock			O'Brien-Fleming		
	Value	$\alpha(s)$	Increment	Value	$\alpha(s)$	Increment
1(0.2)	2.41	0.0079	0.0079	4.56	0.0000	2.6×10^{-6}
2(0.4)	2.41	0.0138	0.0059	3.23	0.0006	0.000574
3(0.6)	2.41	0.0183	0.0045	2.63	0.0045	0.0039
4(0.8)	2.41	0.0219	0.0036	2.28	0.0125	0.0080
5(1.0)	2.41	0.0250	0.0031	2.04	0.0250	0.0125

When 40% of information of the trial is accumulated, with respect to the O'Brien-Fleming boundaries $Z(0.2) = 4.56$ and $z(0.4) = 3.23$, the allowable probability of type I error is only 0.0006. However, the spendable false positive rate is 0.0039 when the third interim analysis is performed at $s = 0.6$. According to Table 10.6.6 in the O'Brien-Fleming group sequential procedure spends very little probability of type I error for the interim analyses performed early in the trial. On the other hand, successive increments of the false positive rate in the Pocock group sequential method is a decreasing function of information as illustrated in Table 10.6.6. Consequently, most of the significance level is spent at an early stage of the trial. Figure 10.6.3 graphically compares the alpha spending functions of the Pocock and O'Brien-Fleming group sequential procedures.

In a maximum duration trial, as noted above, the total information is a random variable, and it is not observed until the trial ends at the maximum duration T_c . It is not uncommon to revise the maximum information specified in the protocol as demonstrated by some of well-known trials sponsored by the U.S. National Institutes of Health such as BHAT (1982) or CAST

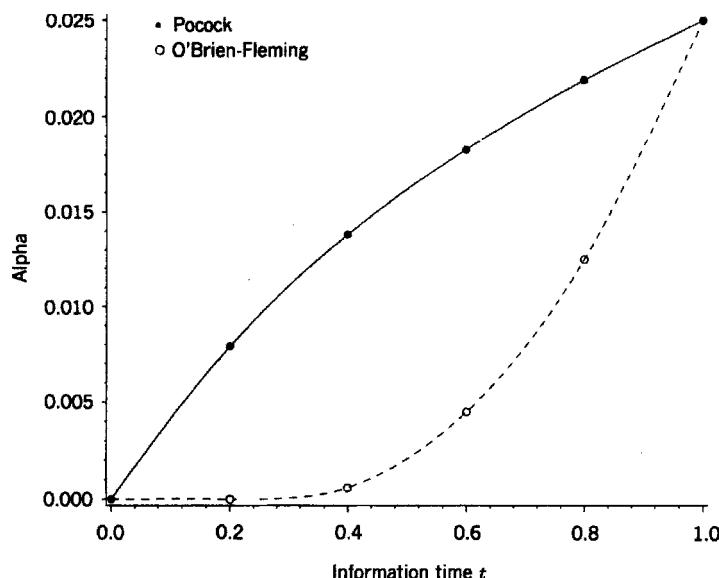


Figure 10.6.3 One-sided 2.5% alpha spending function for Pocock and O'Brien-Fleming boundaries. (Source: Lan and DeMets, 1994.)

(1989). As a result information time cannot be used for computing the boundaries. Lan and DeMets (1983) suggest using a monotone function $g(t)$ to convert the calendar time into information time. However, sometimes $g(t)$ is not known and must be estimated. One such approximation is t/T , the ratio of the current calendar time to the maximum duration. Suppose that the trial has progressed to the calendar time t_k with information accumulated to s_k by previous interim analyses performed at some discrete information time s_i , $i = 1, \dots, k$. Although the total information is not known as t_k , the covariance between $Z(s_i)$ and $Z(s_k)$ can be estimated by $I(t_i)$ to $I(t_k)$, the ratio of information available at calendar time t_i to t_k . Therefore the boundary for the interim analysis at s_k can be computed using the method described above.

The implementation of the alpha spending function requires an advance selection and specification of the spending function in the protocol. One cannot change and choose another spending function in the middle of a trial. Geller (1994) suggests that the spending function be convex and have the property that the same value of a test statistic is more compelling as the sample sizes increase. Since it is flexible and has no requirement for total information and equal increment of information, there is a danger of abuse of the alpha spending function in increasing the frequency of interim analyses as the results approach the boundary. However, DeMets and Lan (1994) reported that altering the frequency of interim analyses has very little impact on the overall significance level if an O'Brien-Fleming-type or a Pocock-type continuous spending function is adopted.

Pawitan and Hallstrom (1990) study the alpha spending function for CAST with the use of a permutation test. A permutation test is conceptually simple, and it provides an exact test for small sample sizes. In addition, it is valid for complicated stratified analysis in which the exact sampling distribution is, in general, unknown and large-sample approximation may not be adequate. Consider the one-sided alternative. For the k th interim analyses, under the assumption of no treatment effect, the null joint permutation distribution of the test statistics (Z_1, \dots, Z_K) can be obtained by random permutation of treatment assignments on the actual data. Let $(Z_{1b}^*, \dots, Z_{Kb}^*)$, $b = 1, \dots, B$, be the statistics computed from B treatment assignments and B be the total number of possible permutations. Given $\alpha(s_1), \alpha(s_2) - \alpha(s_1), \dots, \alpha(s_k) - \alpha(s_{k-1})$, the probabilities of type I error allowed at successive interim analyses, the one-sided boundaries z_1, \dots, z_K can be determined by the formulas

$$\begin{aligned} \frac{\text{Number of } (Z_1^* > z_1)}{B} &= \alpha(s_1), \\ \frac{\text{Number of } (Z_1^* \leq z_1 \text{ and } Z_2^* > z_2)}{B} &= \alpha(s_2) - \alpha(s_1), \\ &\vdots \\ \frac{\text{Number of } (Z_1^* \leq z_1 \text{ and } Z_2^* \leq z_2, \dots, \text{and } Z_K^* > z_K)}{B} &= \alpha(s_K) - \alpha(s_{K-1}). \end{aligned} \quad (10.6.10)$$

If B is very large, then the above method can be executed by a random sample with a replacement of size B . The α spending function for an overall significance level of 2.5% for the one-sided alternative is given by

$$\alpha(s) = \begin{cases} \frac{\alpha}{2}s & \text{if } s < 1, \\ \alpha & \text{if } s = 1. \end{cases}$$

In other words, CAST uniformly spent 50% of the overall probability of type I error until the end of the trial and allocated the other half for the final analysis. The primary event of interest for CAST was death from arrhythmia or cardiac arrest which was originally expected to be 425 based on a projection of an event rate of 11% over a three-year follow-up period. However, the observed event rate was about 2% for the placebo group in the first year. Therefore the event rate was revised to 300 on the second interim analysis on March 30, 1989. The results of interim analyses and boundaries are given in Table 10.6.7. At the time of the first interim analysis, the observed number of events was 29. Therefore $\alpha(29/425) = 0.0125(29/425) = 0.0009$. The corresponding lower boundary using a random sample with replacement of size 4000 computed from (10.6.9) was -3.18 . Since the observed logrank statistic is -2.82 which is greater than -3.18 , the trial continued. Similarly the boundary for the second interim analysis based on an observed 42 events and a revised total expected 300 events was -3.04 . But the observed logrank test statistic was -3.22 which crossed the boundary. This result coupled with other considerations, the DMC of CAST recommended on April 17, 1989, that the trial be terminated because of the adverse effect of the class IC arrhythmic suppressant agents.

CAST study (Pawitan and Hallstrom, 1990) also applied the stochastic curtailment method (Lan et al., 1982) to compute the conditional power of the trial given the current data for consideration of possible termination of the trial due to lack of benefit if the conditional power is too low. Lan and Wittes (1988) propose the B -value as a tool for monitoring data for the calculation of conditional power under the assumption that the current trend indicated by the observed data at information time s continues. For simplicity we consider the one-sample problem. Let $\Theta = \sqrt{N}\mu$, where N is the maximum information that a trial is planned to accumulate and μ is the population mean. The power of rejecting the null hypothesis at the α significance level for a two-sided test conditioning on the

Table 10.6.7 Summary of the Interim Results Based on the Number of Deaths from Arrhythmia or Cardiac Arrest

	Calendar Times for Interim Analyses	
	9/1/88	3/30/89
Placebo	7(576) ^a	9(725)
Active ^b	22(571)	33(730)
Total	29(1147)	42(1455)
Total Expected	425	300
One-sided α increment	0.0009	0.0011
Observed Logrank	-2.82	-3.22
Lower Bound ^c	-3.18	-3.04

^aNumbers in the parentheses are the number of patients assigned to the treatment.

^bActive treatments include encainide and flecainide.

^cLower bounds were computed from a random sample with replacement of size $b = 4000$.

Source: Modified from Pawitan and Hallstrom (1990).

current trend at information time s is given as

$$\begin{aligned}\phi(\Theta) &= 1 - \Phi \left\{ \frac{[Z(\alpha/2) - B(s)/s]}{\sqrt{1-s}} \right\} \\ &= 1 - \Phi \left\{ \frac{[Z(\alpha/2) - Z(s)/\sqrt{s}]}{\sqrt{1-s}} \right\},\end{aligned}\quad (10.6.11)$$

where the B -value at information s is defined as

$$B(s) = \frac{Z(s)}{\sqrt{s}}.\quad (10.6.12)$$

Lan and Wittes (1988) also consider the application of the B -value to the two-sample Z-test, the two-sample Wilcoxon rank sum test, and the logrank test.

Example 10.6.1 Woman's Health Initiative The Women's Health Initiative (WHI), sponsored by the National Heart, Lung, and Blood Institute (NHLBI), the United States National Institutes of Health (NIH), is a large and complex clinical investigation on defining the risks and benefits of strategies that could potentially reduce the incidence of heart disease, breast and colorectal cancer, and fracture in postmenopausal women (The Women's Health Initiative Group, 1998; Writing Group for the WHI Investigators, 2002). One of the components of the WHI protocol is a randomized, double-blind, placebo-controlled, primary prevention trial on hormone replacement therapy (HRT) that was to examine the effect of estrogen (conjugated equine estrogen 0.625 mg/day) plus progestin (medroxyprogesterone acetate 2.5 mg/day) on overall health in 16,608 healthy postmenopausal women aged 50–79 years with intact uterus recruited by 40 clinical centers in the United States between 1993 and 1998. The overall health defined by the WHI is the long-term benefits and risks, including cardiovascular diseases, invasive breast cancer, stroke, pulmonary embolism, endometrial cancer, colorectal cancer, hip fracture, or death due to other causes. However, this study did not address the short-term risks and benefits for the treatment of menopausal symptoms.

Most of the clinical trials discussed in previous chapters focus on the so-called treatment or therapeutic trials in which subjects who suffered from a certain disease defined in the inclusion/exclusion criteria are enrolled. The goal of treatment trials is to evaluate the therapeutic effect of the intervention on a specific disease with a defined severity. On the other hand, the subjects in the primary prevention trials are ostensibly healthy volunteers. As a result, the intervention in the primary prevention trials can affect several diseases and overall health. However, effects of the intervention evaluated in primary prevention trials may be either beneficial for some diseases or adverse for others. In addition, the magnitudes of incidences and the time course of these potentially beneficial and adverse effects are different for different diseases. However, incidences of occurrence of diseases, mortality rate, or morbidity rate of healthy subjects in primary prevention trials are much lower than are therapeutic trials. Consequently, a much larger sample size is required for primary prevention trials and the study duration of the primary prevention trials is much longer than that of the treatment trials. It is, therefore, unlikely to repeat a large-scale primary prevention trial because of cost and long-term nature. For example, the estrogen plus progestin component of the WHI began to enroll 16,608 subjects in the fall of 1997

with expectation of final analysis in the year 2005 after an average of approximately 8.5 years of follow-up.

Note that interim analyses for possible early termination of trials introduced in this chapter are based on a single outcome with respect to a particular disease, usually the primary efficacy endpoint, for treatment trials. This traditional approach, however, is not appropriate for the primary prevention trials due to the inability to repeat and possible direct applications to public health policy. Freedman et al. (1996) argued that a more comprehensive approach to monitoring the primary prevention trials is necessary. They proposed that in addition to considering the effect of the intervention on primary efficacy endpoints, overall health benefit versus risk considerations be incorporated into formal terminating procedures. In other words, a proper balance between the global assessment of overall health and effects of the intervention on primary efficacy endpoints for specific diseases must be reached for monitoring of primary prevention trials.

The primary efficacy endpoint for the estrogen plus progestin component of the WHI is the incidence of coronary heart disease (CHD) because heart disease is the major cause of morbidity and mortality in US postmenopausal women. Besides, cardioprotective effect of HRT cannot be proven with either observational or epidemiological studies. On the other hand, there is a concern that long-term use of estrogen may increase the risk of breast cancer. Therefore, the incidence of invasive breast cancer was chosen as the primary safety endpoint. However, there are some other competing benefits and risks associated with HRT. For the estrogen plus progestin component of the WHI, they are occurrence of stroke, pulmonary embolism, endometrial cancer, colorectal cancer, hip fracture, or death due to other causes. A weighted global index is devised to assess the effect of the estrogen plus progestin on the overall balance of benefit and risk. Let d_1, \dots, d_8 be the observed differences in proportions between the estrogen plus progestin and placebo for the eight outcomes mentioned above. This global index is the weighted sum of d_1, \dots, d_8 :

$$W = w_1 d_1 + \dots + w_8 d_8$$

where w_1, \dots, w_8 represent the weights for these eight outcomes. Freedman et al. (1996) suggested weighing the occurrence of each disease by the expected proportion of diagnosed persons who will die of that disease within a specific number of years of diagnosis. As a result, the weights are less than 1. The weight for deaths from other causes, however, is always 1. This global index can then be monitored just like any other endpoints.

For primary prevention trials, benefits and risks are not symmetric and a proper balance should be maintained between the global assessment and effects on specific diseases. The estrogen plus progestin component of the WHI adopted a mixed approach to possible early termination, which is summarized below:

1. O'Brien-Fleming boundaries are used for each of eight outcomes and for the global index.
2. Asymmetrical upper and lower boundaries are used: a one-sided $\alpha = 0.025$ upper boundary for benefit; one-sided $\alpha = 0.05$ boundaries for adverse effects. In addition, the adverse-effect boundaries were adjusted using the Bonferroni method for the seven outcomes other than invasive breast cancer (CHD, stroke, pulmonary embolism, endometrial cancer, colorectal cancer, hip fracture, or death due to other causes).

3. The trial stops if
 - a. the lower boundary of the adverse effect is crossed
 - b. the upper boundary for benefit is crossed
 - c. the result based on the global index is supportive at a nominal level of 0.20

The independent Data and Safety Monitoring Board (DSMB) conducted its semiannual formal monitoring in the fall of 1997. On May 31, 2002, based on the tenth interim analysis, the DSMB found that the designed specified weighted log-rank test statistic for invasive breast cancer ($z = -3.19$) crossed the designated lower boundary for adverse effect ($z = -2.32$) and the global index was also supportive ($z = -1.62$). In addition, the interim results also suggested that the additional evidence for some increase in CHD, stroke, and pulmonary embolism outweighed the evidence of benefit for fractures and possible benefit for colon cancer over the averaged 5.2 years of follow-up. Therefore, the DSMB recommended early termination of the estrogen plus progestin component of the WHI. Although concepts and statistical methods for interim analyses are intuitively simple and straightforward, the example from the estrogen plus progestin component of the WHI illustrates the complex and yet delicate nature on application of interim analysis to early stopping of a clinical trial.

10.7 DISCUSSION

The statistical inference for censored data covered in this chapter was restricted to parallel group designs in which the time to occurrence of a well-defined clinical event is independent from patient to patient. On the other hand, in the two-sequence, two-period crossover design discussed in Chapter 5, each patient with chronic conditions received both treatments during the study. The result was two censored responses, conditioned under different treatments, were obtained from the same patient. Since they were observed from the same patient, the two censored clinical endpoints were correlated.

There are many examples of correlated censored endpoints in clinical trials. For example, France et al. (1991) report a study whose objective is to evaluate the efficacy of the monotherapy of atenolol at 50 mg twice a day with that of a combination therapy of atenolol and nifedipine in the treatment of angina pectoris. Although both drugs are widely used in the treatment of angina, their modes of action are quite different. Atenolol is a beta-blocker, while nifedipine is a calcium-channel blocker. In addition the dose of atenolol selected at 50 mg b.i.d. for this study was near the top of the dose-response curve such that any additional increase in dose was unlikely to improve the drug's efficacy. Therefore it is expected that the combined therapy of nifedipine and atenolol might provide a more effective treatment of angina than atenolol alone. The study started with a four-week placebo run-in period on atenolol, followed by a two-sequence, two-period crossover design with a duration of four weeks for each treatment period. There was a total of 106 patients who completed the study. At the end of each period, a treadmill test was performed for each patient according to the standard Bruce protocol. The clinical endpoints for assessment of efficacy of the treatment included the time to 1 mm ST segment depression, time to pain, and total time of exercise as well as the reasons for early stopping on the treadmill test. These responses were censored if the treadmill test stopped before symptoms occurred.

Liu and Chow (1993) give another example of correlated censored data in the assessment of clinical equivalence of an albuterol metered dose inhaler (MDI) for acute bronchospasm

between the test and reference formulations in patients with reversible obstructive airway disease. Because of its intended route of administration, we have negligible plasma levels. As a result clinical endpoints are used for assessment of bioequivalence for MDI products. The clinical endpoints for MDI products recommended by the FDA guidance are derived from the volume of air forced out of the lung within one second (FEV_1) measured at 0, 10, 15, 30, 60, 90, 120, 180, 240, 300, and 360 minutes after dosing. One of the clinical endpoints derived from FEV_1 is the time to the onset of a therapeutic response which is defined as the event that the FEV_1 measurement evaluated within 30 minutes after dosing exceeds 115% of its baseline measurement at time 0. In order to reduce the variability between patients, crossover designs are the designs of choice in evaluations of bioequivalence for MDI products. As a result the censored data, such as the time to the onset of therapeutic response in the same patient, are correlated.

Several methods are available for paired censored clinical endpoints. For example, O'Brien and Fleming (1987) proposed the paired Prentice-Wilcoxon statistic, Huster, Brookmeyer, and Self (1989) suggest parametric models with adjustments by covariates, Holt and Prentice (1974) and Kalbfleisch and Prentice (1980) employ the Cox proportional hazard model to analyze paired censored data. However, these methods fail to take into account the structure of crossover designs. Recently, under assumption of no carryover effect, France et al. (1991) and Liu and Chow (1993) extended the method of Cox's proportional hazard model for paired censored data proposed by Kalbfleisch and Prentice (1980) to the crossover design.

Basically, the method proposed by Kalbfleisch and Prentice (1980) is the sign test generalized to the censored data, which can be derived from the binary logistic regression discussed in Chapter 9. For each sequence within a two-sequence, two-period design, the treatment effect can be estimated as the regression coefficient for the treatment indicator variable according to the method suggested by Kalbfleisch and Prentice (1980). The overall estimated treatment effect is then the average of individual regression coefficients estimated from each sequence. On the other hand, the period effect is the difference in individual regression coefficients between the two sequences. However, under Cox's proportional hazard model, it is not possible to provide a test for the presence of the carryover effect. Liu and Chow (1993) suggest using the hazard ratio as a criterion in the assessment of bioequivalence between MDI products. Jung and Su (1995) propose a nonparametric estimation procedure for the difference or ratio of median failure times for the correlated censored observations from a two-sequence, two-period crossover study. They recommended the use of a 95% confidence interval for the difference or ratio of median failure time as a criterion in the assessment of equivalence between drug products.

As was mentioned above, the method for correlated censored data obtained from crossover designs proposed by France et al., (1991) and Liu and Chow (1993) is an extension of a sign test that may not be powerful under certain alternatives. Feingold and Gillespie (1996) therefore suggest two methods to test the treatment effect. The first method is to transform the data into a score and to analyze the score according to the standard procedure for continuous data under the linear model for a two-sequence, two-period crossover design, assuming compound symmetry in the covariance matrix. They recommended the use of the Gehan score (1965b). For a two-sequence, two-period crossover design, the treatment effect, period effect, and carryover (or sequence) effect can be obtained from within-patient contrasts. The second testing method, based on the censoring pattern in each period and signs of linear contrast, is to classify the result of each patient as uncensored, left-censored, right-censored, or undefined. Feingold and Gillespie (1996) then suggest testing the differences

between the two sequences with a distribution-free test developed by Gehan (1965a) and Mantel (1967), which is an extension of the two-sample generalized Wilcoxon test for right-censored data. Feingold and Gillespie (1996) also select the average quantile as the parameter for estimation of the treatment effect. The average quantile is referred to as the average distance of each survival curve from the origin, computed over the quantile range where all of the survival curves are defined. If all data are uncensored, then the average quantile is the mean. However, their procedure for hypothesis testing is not in accordance with the estimation results. In other words, if one of the proposed tests indicates a statistically significant treatment effect, this does not imply that the confidence interval for that parameter does not include zero.

The group sequential procedures for interim analyses are basically in the context of hypothesis testing which is aimed at pragmatic study objectives concerning which treatment is better. However, most new treatments such as cancer drugs are very expensive, very toxic, or both. As a result only if the degree of the benefit provided by the new treatment exceeds some minimum clinically significant requirement, it will then be considered for the treatment of the intended medical conditions. Therefore an adequate well-controlled trial should be able to provide not only the qualitative evidence whether the experimental treatment is effective but also the quantitative evidence on the unbiased estimation of the size of the effectiveness or safety over the placebo given by the experimental therapy. For a fixed sample design without interim analyses for early termination, it is possible to achieve both qualitative and quantitative goals with respect to the treatment effect. However, with the group sequential procedure the beneficial size of the experimental treatment by the maximum likelihood method is usually overestimated because of the choice of stopping rule. Jennison and Turnbull (1990) point out that the sample mean might not be even contained in the final confidence interval. As a result, estimation of the size of treatment effect has received a lot of attention. Various estimation procedures have been proposed such as the modified maximum likelihood estimator (MLE), the median unbiased estimator (MUE) and the midpoint of the equal two tailed 90% confidence interval. For more detail, see Cox (1952), Tsiatis et al. (1984), Kim and DeMets (1987), Kim (1989), Chang and O'Brien (1986), Chang et al. (1989), Chang (1989), Hughes and Pocock (1988), and Pocock and Hughes (1989).

The estimation procedures proposed in the above literature require extensive computation. On the other hand, simulation results (Kim 1989; Hughes and Pocock 1989) show that the alpha spending function corresponding to the O'Brien-Fleming group sequential procedure is quite concave and allocates only a very small total nominal significance level to early stages of the interim analyses, and hence the bias, variance, and mean square error of the point estimator following O'Brien-Fleming procedure are also the smallest. Current research has mainly focused on the estimation of the size of the treatment effect for the primary clinical endpoints on which the group sequential procedure is based. However, there are many other secondary efficacy and safety endpoints to be evaluated in the same trial. The impact of early termination of a trial based on the results from primary clinical endpoints on the statistical inference for these secondary clinical endpoints is unclear. In addition, group sequential methods and their followed estimation procedures so far are only concentrated on the population average. On the other hand, inference of variability is sometimes also of vital importance for certain classes of drug products and diseases. Research on estimation of variability following early termination is still lacking. Other areas of interest for interim analyses include clinical trials with more than two treatments and bioequivalence assessment. For group sequential procedures for trials with multiple treatments, see Hughes (1993) and Proschan et al. (1994). For the group sequential bioequivalence testing procedure, see Gould (1995b).

11

SAMPLE SIZE DETERMINATION

11.1 INTRODUCTION

A major consideration of most clinical studies is to determine whether the drug under investigation is effective and safe. During the planning stage of a clinical study, the following questions are of particular interest to the investigators: (1) How many subjects are needed in order to have a desired power (e.g., 80% chance of correctly detecting a clinically meaningful difference? (2) What's the *trade-off* if only a small number of subjects are available for the study due to limited budget and/or some medical considerations? To address these questions, a statistical evaluation for sample size determination/justification is often employed. Sample size determination usually involves the calculation of a required sample size for some desired statistical properties such as precision and power, whereas sample size justification provides statistical justification for a selected sample size which may be small in number because of budget constraints or some medical considerations. In this chapter our emphasis will be placed on sample size determination. The concept can be easily implemented to provide statistical justification for a selected sample size.

For a given study, sample size is usually determined based on some criteria on type I and/or type II errors. For example, we can choose sample size in such a way that there is a desired precision at a fixed confidence level (or fixed type I error). This approach is referred to as precision analysis for sample size determination. The disadvantage of the precision analysis is that it has a small chance of detecting a true difference. As an alternative, a pre-study power analysis is usually conducted to calculate sample size. The concept of power analysis is to select an acceptable type II error for a fixed type I error. In other words, the selected sample size will have a desired power for correctly detecting a clinical/scientific meaningful difference at a fixed type I error rate. In most clinical trials the pre-study power

analysis for sample size determination is the most commonly used method for choosing the sample size. Therefore, in this chapter we will focus on sample size determination based on power analysis for various situations of clinical trials.

To perform a pre-study power analysis for sample size determination, the power function of an appropriate test for the hypotheses of interest is necessarily characterized. The hypotheses should be established to reflect the study objectives under the study design. In practice, it is not uncommon to observe discrepancies among study objective (hypotheses), study design, statistical analysis (test statistic), and sample size calculation. These inconsistencies often result in (1) wrong test for right hypotheses, (2) right test for wrong hypotheses, (3) wrong test for wrong hypotheses, or (4) right test for right hypotheses with insufficient power. Therefore, before the sample size can be determined, it is suggested that the following be carefully considered: (1) the study objectives or the hypotheses of interest be clearly stated, (2) a valid design with appropriate statistical tests be used, and (3) the sample size be determined based on a test for the hypotheses of interest.

In the next section we will introduce some basic concepts for sample size determination and outline the precision analysis approach and the method of pre-study power and analysis for sample size calculation. The pre-study power analysis for sample size determination will be considered to derive formulas for complex clinical trials with different study objectives. In Section 11.3 and 11.4 sample size calculations for two samples comparing two treatments and multiple samples comparing more than two treatments are discussed, respectively. The sample size required for complex clinical trials with survival endpoints is outlined in Section 11.5. In Section 11.6 we explore sample size determination for dose-response studies. Sample size calculation under the structure of crossover designs is described in Section 11.7. In Section 11.8, some recent developments for sample size required for equivalence or noninferiority trials with binary or censored data as primary endpoints are reviewed. Section 11.9 gives tables for sample size determination of optimal two-stage designs for phase II cancer trials. The final section provides formulas for sample size calculation for comparing variabilities in clinical research.

11.2 BASIC CONCEPT

Study Objectives and Hypotheses

In most clinical trials the primary study objective is related to the evaluation of the effectiveness and safety of a drug product. For example, it may be of interest to show that the study drug is effective and safe compared to a placebo for some intended indications. In some cases it may be of interest to show that the study drug is as effective as, superior to, or equivalent to an active control agent or a standard therapy. In practice, hypotheses regarding medical or scientific questions of the study drug are usually formulated based on the primary study objectives. The hypotheses are then evaluated using appropriate statistical tests under a valid study design. To ensure that the test results will have the desired power with a certain degree of accuracy and reliability, a sufficient number of subjects is needed. As a result sample size determination plays a crucial role in the design of clinical trials.

For a given study objective, a corresponding hypothesis can be formulated. As was discussed in Section 2.6, in clinical trials a hypothesis is usually referred to as a postulation, assumption, or statement that is made about the population regarding the effectiveness and safety of a drug under investigation. For example, the statement that there is a direct drug effect is a hypothesis regarding the treatment effect. For testing the hypotheses of interest,

a random sample is usually drawn from the targeted population to evaluate hypotheses about the drug product. A statistical test is then performed to determine whether the null hypothesis should be rejected at a prespecified significance level. Based on the test result, conclusion(s) regarding the hypotheses can be drawn.

The selection of hypotheses depends on the study objectives. For example, if we want to demonstrate the effectiveness of a drug product compared to a placebo in terms of an outcome variable, we could consider the following hypothesis:

$$\begin{aligned} H_0: \mu_T &= \mu_P, \\ \text{vs. } H_a: \mu_T &\neq \mu_P, \end{aligned} \quad (11.2.1)$$

where μ_T and μ_P are the mean response of the outcome variable for the test drug and the placebo, respectively. As was discussed before, a typical approach is to show that there is a statistically significant difference between the test drug and the placebo by rejecting the null hypothesis and then demonstrating that there is a high chance of correctly detecting a clinically meaningful difference if such a difference truly exists.

On the other hand, one may want to show that the test drug is as effective as an active agent or a standard therapy. In this case Blackwelder (1982) suggests testing the following hypotheses:

$$\begin{aligned} H_0: \mu_T - \mu_S &\leq -\delta, \\ \text{vs. } H_a: \mu_T - \mu_S &> -\delta, \end{aligned} \quad (11.2.2)$$

where μ_S is the mean for a standard therapy and δ is a difference of clinical importance. The concept is to reject the null hypothesis and conclude that the difference between the test drug and the standard therapy is less than a clinically meaningful difference δ and hence that the test drug is as effective as the standard therapy. This study objective is not uncommon in clinical trials especially when the test drug is considered to be less toxic, easier to administer, or less expensive than the established standard therapy.

To show superiority of a test drug over an active control agent or a standard therapy, we may consider the following hypotheses:

$$\begin{aligned} H_0: \mu_T - \mu_S &\leq \delta, \\ \text{vs. } H_a: \mu_T - \mu_S &> \delta, \end{aligned} \quad (11.2.3)$$

The rejection of the above null hypothesis suggests that the difference between the test drug and the standard therapy is greater than a clinically meaningful difference. Therefore, we can conclude that the test drug is superior to the standard therapy by rejecting the null hypothesis of (11.2.3).

In practice, unless there is some prior knowledge regarding the test drug, usually we do not know the performance of the test drug as compared to the standard therapy. Therefore hypotheses (11.2.2) and (11.2.3) are not preferred because they have predetermined the performance of the test drug as compared to the standard therapy. As an alternative, the following hypotheses for therapeutic equivalence are usually considered:

$$\begin{aligned} H_0: |\mu_T - \mu_S| &\geq \delta, \\ \text{vs. } H_a: |\mu_T - \mu_S| &< \delta. \end{aligned} \quad (11.2.4)$$

We then conclude that the difference between the test drug and the standard therapy is of no clinical importance if the null hypothesis of (11.2.4) is rejected.

It should be noted that a valid sample size calculation can only be done based on appropriate statistical tests for the hypotheses that can reflect the objectives under a valid study design. It is then suggested that the hypotheses be clearly stated when performing a sample size calculation. Each of the above hypotheses has a different requirement for the sample size in order to achieve a desired power or precision of the corresponding tests.

Type I and Type II Errors

In practice, two kinds of errors occur when testing hypotheses. If the null hypothesis is rejected when it is true, then a type I error has occurred. If the null hypothesis is not rejected when it is false, then a type II error has been made. The probabilities of making type I and type II errors, denoted by α and β , respectively, are given below:

$$\begin{aligned}\alpha &= P\{\text{type I error}\} \\ &= P\{\text{reject } H_0 \text{ when } H_0 \text{ is true}\}, \\ \beta &= P\{\text{type II error}\} \\ &= P\{\text{fail to reject } H_0 \text{ when } H_0 \text{ is false}\}.\end{aligned}$$

The probability of making a type I error, α , is called the level of significance. The power of the test is defined as the probability of correctly rejecting the null hypothesis when the null hypothesis is false, namely,

$$\begin{aligned}\text{Power} &= 1 - \beta \\ &= P\{\text{reject } H_0 \text{ when } H_0 \text{ is false}\}.\end{aligned}$$

For example, suppose that we want to test the following hypotheses:

$$\begin{aligned}H_0: &\text{The drug is ineffective;} \\ \text{vs. } H_a: &\text{The drug is effective.}\end{aligned}$$

Then a type I error occurs if we conclude that the drug is effective when in fact it is not. On the other hand, a type II error occurs if we claim that the drug is ineffective when in fact it is effective. In clinical trials none of these errors is desirable. A typical approach is to avoid type I error but at the same time to decrease type II error so that there is a high chance of correctly detecting a drug effect when the drug is indeed effective. Chow and Liu (2000) illustrate the relationship between type I error and type II error (or power). It appears that α decreases as β increases and α increases as β decreases. The only way of decreasing both α and β is to increase the sample size.

Typically, sample size can be determined by controlling either a type I error (or confidence level) or a type II error (or power). In what follows we will introduce two concepts for sample size determination, namely precision analysis based on type I error and power analysis based on type II error.

Precision Analysis

In practice, the maximum probability of committing a type I error that one can tolerate is usually considered the level of significance. The confidence level, $1 - \alpha$, then reflects the probability or confidence of not rejecting the true null hypothesis. Since the confidence

interval approach is equivalent to the method of hypotheses testing, we can determine the sample size required based on the type I error rate using the confidence interval approach. For a $(1 - \alpha)100\%$ confidence interval, the precision of the interval depends on its width. The narrower the interval is, the more precise the inference is. Therefore, the precision analysis for sample size determination is to consider the maximum half width of the $(1 - \alpha)100\%$ confidence interval of the unknown parameter that one is willing to accept. Note that the maximum half width of the confidence interval is usually referred to as the *maximum error* of an estimate of the unknown parameter. For example, let Y_1, Y_2, \dots, Y_n be independent and identically distributed normal random variables with mean μ and variance σ^2 . When σ^2 is known, a $(1 - \alpha)100\%$ confidence interval for μ can be obtained as

$$\bar{Y} \pm Z(\alpha/2) \sigma/\sqrt{n},$$

where $Z(\alpha/2)$ is the upper $(\alpha/2)$ th quantile of the standard normal distribution. The maximum error, denoted by E in estimating the value of μ that one is willing to accept, is then defined as

$$E = |\bar{Y} - \mu| = Z(\alpha/2) \sigma/\sqrt{n}.$$

Thus the sample size required can be chosen as

$$n = \frac{Z(\alpha/2)^2 \sigma^2}{E^2}. \quad (11.2.5)$$

Note that the maximum error approach for choosing n is to attain a specified precision while estimating μ which is derived only based on the interest of type I error. A nonparametric approach can be obtained by using the following Chebyshev inequality:

$$P\{|\bar{Y} - \mu| \leq E\} \geq 1 - \frac{\sigma^2}{nE^2}$$

and hence

$$n = \frac{\sigma^2}{\alpha E^2} \quad (11.2.6)$$

Note that the precision analysis for sample size determination is very easy to apply based on either (11.2.5) or (11.2.6). For example, suppose that we wish to have a 95% assurance that the error in the estimated mean is less than 10% of the standard deviation (i.e., 0.1σ). Thus

$$Z(\alpha/2)\sigma/\sqrt{n} = 0.1\sigma.$$

Hence

$$n = \frac{Z(\alpha/2)^2 \sigma^2}{E^2} = \frac{(1.96)^2 \sigma^2}{(0.1\sigma)^2} = 384.2 \approx 385.$$

The above concept can be applied to binary data (or proportions). In addition, it can be easily implemented for sample size determination when comparing two treatments. Table 11.2.1 provides a summary for sample size determination based on precision analysis for situations where there are one and two samples, respectively.

Table 11.2.1 Sample Size Determination Based on Precision Analysis

Parameter	Statistic	Confidence Interval	Sample Size
μ	\bar{Y}	$\bar{Y} \pm Z(\alpha/2) \frac{\sigma}{\sqrt{n}}$	$n = \frac{Z(\alpha/2)^2 \sigma^2}{E^2}$
$\mu_1 - \mu_2$	$\bar{Y}_1 - \bar{Y}_2$	$(\bar{Y}_1 - \bar{Y}_2) \pm Z(\alpha/2) \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}}$	$n = \frac{Z(\alpha/2)^2 (\sigma_1^2 + \sigma_2^2)}{E^2}$

Power Analysis

Since a type I error is usually considered to be a more important and/or serious error than one would like to avoid, a typical approach in hypothesis testing is to control α at an acceptable level and try to minimize β by choosing an appropriate sample size. In other words, the null hypothesis can be tested at a predetermined level (or nominal level) of significance with a desired power. This concept for determination of the sample size is usually referred to as a *power analysis* for sample size determination.

For determination of the sample size based on a power analysis, the investigator(s) is required to specify the following information: First of all, a significance level has to be selected at which the chance of wrongly concluding that a difference exists when in fact there is no real difference (type I error) can be tolerated. Typically, a 5% level of significance is chosen to reflect at 95% confidence regarding the unknown parameter. Second, the desired power has to be selected at which the chance of correctly detecting a difference when the difference truly exists that one wishes to achieve. A conventional choice of power is either 90% or 80%. Third, a clinically meaningful difference must be specified. In most clinical trials the objective is to demonstrate that effectiveness and safety of a drug under study compared to a placebo. Therefore, it is important to specify what difference in terms of the primary endpoint is considered clinically or scientifically important. We denote such a difference by Δ . If the investigator will settle for detecting only a large difference, then fewer subjects will be needed. If the difference is relatively small, a larger study group (i.e., a large number of subjects) will be needed. Finally, the knowledge regarding the standard deviation (i.e., σ) of the primary endpoint considered in the study is required for the sample size determination. A very precise method of measurement (i.e., a small σ) will permit detection of any given difference with a much smaller sample size than would be required with a less precise measurement.

In the following sections, we describe some methods for sample size determination based on the power analyses for various situations that are commonly encountered in clinical trials.

11.3 TWO SAMPLES

One-Sample Test for Mean

Suppose that one wishes to test the following hypotheses:

$$\begin{aligned} H_0: \mu &= \mu_0, \\ \text{vs. } H_a: \mu &> \mu_0, \end{aligned} \tag{11.3.1}$$

with a significance level α when the variance σ^2 is known. For a specific alternative hypothesis, say

$$H_a: \mu = \mu_0 + \Delta,$$

where $\Delta > 0$ is a constant, the power of the test is given by

$$\begin{aligned} 1 - \beta &= P\{\text{reject } H_0 | H_a \text{ is true}\} \\ &= P\left\{\frac{\bar{Y} - (\mu_0 + \Delta)}{\sigma/\sqrt{n}} > Z(\alpha) - \frac{\Delta}{\sigma/\sqrt{n}} \mid \mu = \mu_0 + \Delta\right\}. \end{aligned}$$

Under the alternative hypothesis that $\mu = \mu_0 + \Delta$, the test statistic

$$\frac{\bar{Y} - (\mu_0 + \Delta)}{\sigma/\sqrt{n}}$$

follows a standard normal variable. Therefore

$$1 - \beta = P\left\{Z > Z(\alpha) - \frac{\Delta\sqrt{n}}{\sigma}\right\},$$

from which we conclude that

$$-Z(\beta) = Z(\alpha) - \frac{\Delta\sqrt{n}}{\sigma}$$

and hence

$$n = \frac{\sigma^2[Z(\alpha) + Z(\beta)]^2}{\Delta^2}. \quad (11.3.2)$$

This result is also true when the alternative is

$$H_a: \mu < \mu_0.$$

Then, for testing a two-sided hypotheses

$$\begin{aligned} H_0: \mu &= \mu_0, \\ \text{vs. } H_a: \mu &\neq \mu_0, \end{aligned}$$

we obtain $1 - \beta$ power for a specified alternative when

$$n = \frac{\sigma^2[Z(\alpha/2) + Z(\beta)]^2}{\Delta^2} \quad (11.3.3)$$

For the one-sample test for a proportion, the required sample size can be similarly derived. Let Y be the Bernoulli random variable of interest with the probability of success P and the probability of failure $1 - P$. The objective of the study is to choose between $H_0: P = P_0$ and $H_a: P = P_1 (P_1 > P_0)$ based on a sample of size n . The sample proportion

$$p = \frac{1}{n} \sum_{i=1}^n Y_i$$

approximately follows a normal distribution with mean P and variance $P(1 - P)/n$, then the required sample size is given by

$$n = \frac{[Z(\alpha)\sqrt{P_0(1 - P_0)} + Z(\beta)\sqrt{P_1(1 - P_1)}]^2}{(P_1 - P_0)^2} \quad (11.3.4)$$

To simplify (11.3.4), many authors (e.g., Lemeshow, Hosmer, and Stewart, 1981; Haseman, 1978) have suggested to consider the following arcsin transformation:

$$A(p) = 2 \arcsin \sqrt{p}$$

before performing the normal test. In this case (11.3.4) becomes

$$n = \frac{[Z(\alpha) + Z(\beta)]^2}{[A(P_1) - A(P_0)]^2} \quad (11.3.5)$$

The above result is also valid for the case where $H_a: P = P_1$ ($P_1 < P_0$). Hence, for testing a two-sided hypotheses, the required sample size is given by

$$n = \frac{[Z(\alpha/2) + Z(\beta)]^2}{[A(P_1) - A(P_0)]^2}. \quad (11.3.6)$$

Two-Sample Test for Comparing Means

A similar procedure can be applied to determine the sample size required for achieving a specific power in comparing two treatment means. Let μ_1 and μ_2 denote the means for treatment 1 and treatment 2, respectively. The hypotheses of interest are then given by

$$\begin{aligned} H_0: \mu_1 &= \mu_2, \\ \text{vs. } H_a: \mu_1 &\neq \mu_2. \end{aligned} \quad (11.3.7)$$

Assuming that σ_1^2 and σ_2^2 are known, for a specific alternative hypothesis that $\mu_1 = \mu_2 + \Delta$, the power is given by

$$1 - \beta = P\left\{ \left| \frac{\bar{Y}_1 - \bar{Y}_2}{\sigma_d} \right| > Z(\alpha/2) | \mu_1 = \mu_2 + \Delta \right\},$$

where

$$\sigma_d = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Therefore

$$\beta = P\left\{ -Z(\alpha/2) - \frac{\Delta}{\sigma_d} < \frac{(\bar{Y}_1 - \bar{Y}_2) - \Delta}{\sigma_d} < Z(\alpha/2) - \frac{\Delta}{\sigma_d} | \mu_1 = \mu_2 + \Delta \right\}.$$

Under the alternative hypothesis, the statistic

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - \Delta}{\sigma_d}$$

is a standard normal variable. Therefore

$$\beta = P \left\{ -Z(\alpha/2) - \frac{\Delta}{\sigma_d} < Z < Z(\alpha/2) - \frac{\Delta}{\sigma_d} \right\},$$

from which we conclude that

$$-Z(\beta) = Z(\alpha/2) - \frac{\Delta}{\sigma_d}.$$

If we assume that $n = n_1 = n_2$, then

$$\begin{aligned} n &= \frac{(\sigma_1^2 + \sigma_2^2)[Z(\alpha/2) + Z(\beta)]^2}{\Delta^2} \\ &= \frac{2\sigma^2[Z(\alpha/2) + Z(\beta)]^2}{\Delta^2}, \quad \text{if } \sigma_1^2 = \sigma_2^2. \end{aligned} \quad (11.3.8)$$

For the one-sided test, the above expression for the required sample size becomes

$$n = \frac{(\sigma_1^2 + \sigma_2^2)[Z(\alpha) + Z(\beta)]^2}{\Delta^2}. \quad (11.3.9)$$

Note that when the population variance is unknown, the choice of sample size is not straightforward. For example, in testing the null hypothesis of (11.3.1), when the true value is $\mu = \mu_0 + \Delta$, the statistic

$$\frac{\bar{Y} - (\mu_0 + \Delta)}{s/\sqrt{n}}$$

follows a noncentral t distribution and noncentrality parameter $\delta = \Delta/\sigma$. Tables 11.3.1 and 11.3.2 provide sample sizes for the t test of the mean and the difference between treatments, respectively, for various value of δ .

When the outcome variable is dichotomous (e.g., either improves or does not improve), the outcome variable of interest is the proportion of patients who have the disease rather than the mean of a specified measurement. Let P_1 and P_2 be the proportions of success (e.g., cure or improvement) in the treatment group and the control group, respectively. Then the sample size can be similarly determined based on two-sided test as follows:

$$n = \frac{[Z(\alpha/2)\sqrt{2P(1-P)} + Z(\beta)\sqrt{P_1(1-P_1) + P_2(1-P_2)}]^2}{(P_1 - P_2)^2}, \quad (11.3.10)$$

where $P = 1/2(P_1 + P_2)$. An arcsin transformation, as defined earlier, gives

$$n = \frac{[Z(\alpha/2) + Z(\beta)]^2}{[A(P_1) - A(P_2)]^2}. \quad (11.3.11)$$

Stolley and Strom (1986) indicate that (11.3.7) is useful for prospective cohort studies and retrospective case-control studies as well. For prospective cohort studies, P_1 and P_2 represent the smallest proportion developing the disease in the exposed study group and the proportion expected to develop the disease in the unexposed control group, respectively. The ratio P_1/P_2 ,

Table 11.3.1 Sample Size for the t Test of the Mean

Value of $\Delta = \frac{\mu - \mu_0}{\sigma}$	Level of t Test									
	$\alpha = 0.005$					$\alpha = 0.01$				
	$\alpha = 0.01$	$\alpha = 0.02$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$
One-Sided Test	0.15	110	134	110	109	115	139	119	128	99
Two-Sided Test	0.20	78	99	85	85	85	90	90	64	122
$\beta =$	0.25	45	58	66	66	66	63	67	34	97
	0.30	37	101	81	81	81	37	109	88	101
	0.35	115	97	77	77	75	30	93	84	70
	0.40	92	77	62	62	63	30	67	51	55
	0.45	125	125	51	51	55	43	76	54	41
	0.50	100	75	63	63	66	25	76	44	21
	0.55	83	63	53	42	55	46	45	37	28
	0.60	71	53	45	36	46	36	63	37	15
	0.65	61	46	39	31	47	39	31	38	32
	0.70	53	40	34	28	47	35	30	24	18
	0.75	47	36	30	25	42	31	27	21	16
	0.80	41	32	27	22	37	28	24	19	12
	0.85	37	29	24	20	33	25	21	17	13
	0.90	34	26	22	18	29	23	19	16	12
	0.95	31	24	20	17	11	27	21	18	14
	1.00	28	22	19	16	10	25	19	13	10
	1.1	24	19	16	14	9	21	16	13	9
	1.2	21	16	14	12	8	18	14	12	7
	1.3	18	15	13	11	8	16	13	11	6
	1.4	16	13	12	10	7	14	11	10	5
	1.5	15	12	11	9	7	14	10	9	6

Table 11.3.1 (Continued)

		Level of t Test															
		$\alpha=0.01$					$\alpha=0.025$					$\alpha=0.05$					
		$\alpha=0.005$		$\alpha=0.01$		$\alpha=0.02$		$\alpha=0.05$		$\alpha=0.025$		$\alpha=0.05$		$\alpha=0.1$			
One-Sided Test	Two-Sided Test	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	
$\beta =$	$\frac{\text{Value of } \Delta = \frac{\mu - \mu_0}{\sigma}}{\sigma}$	1.6	13	11	10	8	6	12	10	9	7	5	10	8	7	6	5
		1.7	12	10	9	8	6	11	9	8	7	5	9	7	6	5	5
		1.8	12	10	9	8	6	10	8	7	7	5	8	7	6	7	6
		1.9	11	9	8	7	6	10	8	7	6	5	8	6	6	7	5
		2.0	10	8	8	7	5	9	7	7	6	5	7	6	5	6	6
		2.1	10	8	7	7	5	8	7	6	6	5	7	6	5	6	6
		2.2	9	8	7	6	5	8	7	6	5	5	7	6	5	6	5
		2.3	9	7	7	6	5	8	6	6	5	5	6	5	5	6	5
		2.4	8	7	7	6	5	7	6	6	5	5	6	5	6	6	6
		2.5	8	7	6	6	5	7	6	6	5	5	6	5	6	6	6
		3.0	7	6	6	5	5	6	5	5	5	5	5	5	5	5	5
		3.5	6	5	5	5	5	6	5	5	5	5	5	5	5	5	5
		4.0	6														

Table 11.3.2 Sample Size for the t Test of the Difference Between Two Means

$\beta = \frac{\mu_1 - \mu_2}{\sigma}$	One-Sided Test	$\alpha=0.005$						$\alpha=0.01$						$\alpha=0.025$						$\alpha=0.05$							
		$\alpha=0.01$			$\alpha=0.02$			$\alpha=0.01$			$\alpha=0.05$			$\alpha=0.05$			$\alpha=0.01$			$\alpha=0.05$			$\alpha=0.1$				
		0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	
0.20		124																									
0.25		123																									
0.30		123																									
0.35		123																									
0.40		123																									
0.45		123																									
0.50		123																									
0.55		123																									
0.60		123																									
0.65		123																									
0.70		123																									
0.75		123																									
0.80		123																									
0.85		123																									
0.90		123																									
0.95		123																									
1.00		123																									
1.1		123																									
1.2		123																									
1.3		123																									
1.4		123																									
1.5		123																									
1.6		123																									

Table 11.3.2 (Continued)

		Level of <i>t</i> Test																		
One-Sided Test	Two-Sided Test	$\alpha = 0.005$				$\alpha = 0.01$				$\alpha = 0.025$				$\alpha = 0.05$						
		$\alpha = 0.01$	$\alpha = 0.02$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$				
$\beta =$	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5	0.01	0.05	0.1	0.2	0.5
Value of $\Delta = \frac{\mu_1 - \mu_2}{\sigma}$																				
1.7	19	15	15	10	7	17	13	11	9	6	14	11	9	7	4	12	9	7	6	3
1.8	17	13	11	10	6	15	12	10	8	5	13	10	8	6	4	11	8	7	5	5
1.9	16	12	11	9	6	14	11	9	8	5	12	9	7	6	4	10	7	6	5	5
2.0	14	11	10	8	6	13	10	9	7	5	11	8	7	6	4	9	7	6	4	4
2.1	13	10	9	8	5	12	9	8	7	5	10	8	6	5	3	8	6	5	4	4
2.2	12	10	8	7	5	11	9	7	6	4	9	7	6	5	3	8	6	5	4	4
2.3	11	9	8	7	5	10	8	7	6	4	9	7	6	5	3	7	5	5	4	4
2.4	11	9	8	6	5	10	8	7	6	4	8	6	5	4	3	7	5	4	4	4
2.5	10	8	7	6	4	9	7	6	5	4	8	6	5	4	3	6	5	4	3	3
3.0	8	6	6	5	4	7	6	5	4	3	6	5	4	4	3	5	4	3	4	3
3.5	6	5	5	4	3	6	5	4	4	3	5	4	4	3	2	4	3	4	3	4
4.0	6	5	4	4		5	4	4	3		4	4	3	2	1	4	3	4	3	4

which is the incidence rate in the exposed group divided by the incidence rate in the control group, is usually referred to as the relative risk. A relative risk greater than 1.0 indicates that the exposure appears to increase the risk of the outcome. A relative risk less than 1.0 indicates that the exposure appears to decrease the risk of the outcome. A relative risk of 1.0 indicates that there is no association between the exposure and the outcome. For retrospective case-control studies, P_1 and P_2 represent the smallest proportion exposed to the risk factor of interest that one would consider important to detect in the diseased (case) group and the proportion expected to experience the exposure of interest in the undiseased (control) group.

A comprehensive review of the formulas and tables for the determination of sample sizes and power in clinical trials for testing differences in proportions for the two-sample design can be found in Sahai and Khurshid (1996), who provide separate formulas for equal and unequal treatment group sizes, formulas for the calculation of power given the sample size, and complete references for all formulas and tables cited.

Example 11.3.1 Suppose that a pharmaceutical company is interested in conducting a clinical trial to compare two cholesterol lowering agents for treatment of hypercholesterolemic patients. The primary efficacy parameter is a low-density lipidprotein cholesterol (LDL-C). Suppose that a difference of 8% in the percent change of LDL-C is considered a clinically meaningful difference and that the standard deviation is assumed to be 15%. Then, by (11.3.8), at $\alpha = 0.05$, the required sample size for having an 80% power can be obtained as follows:

$$\begin{aligned} n &= \frac{2\sigma^2[Z(\alpha/2) + Z(\beta)]^2}{\Delta^2} \\ &= \frac{2(15)^2[1.96 + 0.842]^2}{(8)^2} \\ &= 55.2 \approx 56. \end{aligned}$$

Therefore a sample size of 56 patients per arm is required to obtain an 80% power for detection of an 8% difference in percent change of LDL-C for the intended clinical study.

Example 11.3.2 A pharmaceutical company is interested in examining the effect of an antidepressant agent in patients with generalized anxiety disorder. A double-blind, two-arm parallel, placebo-controlled randomized trial is planned. To determine the required sample size for achieving an 80% power, the HAM-A scores is considered as the primary efficacy variable. It is believed that a difference of 4 in the HAM-A scores between the antidepressant and the placebo is of clinical importance. Assuming that the standard deviation is 7.0 obtained from previous studies, the required sample size can be obtained based on (11.3.8), which is given by

$$\begin{aligned} n &= \frac{2\sigma^2[Z(\alpha/2) + Z(\beta)]^2}{\Delta^2} \\ &= \frac{2(7)^2[1.96 + 0.842]^2}{(4)^2} \\ &= 48.1 \approx 49. \end{aligned}$$

Therefore a sample size of 49 patients per arm is required to obtain an 80% power for detection of a difference of 4 in HAM-A scores between the antidepressant agent and the placebo when performing a two-tailed test at $\alpha = 0.05$.

Example 11.3.3 To evaluate the efficacy and safety of an anti-infective agent compared to active control in the treatment of lower respiratory tract infections, a clinical trial is planned. A response rate of 90% for the active control agent is assumed (based on previous studies). In the interest of having an 80% power for detection of a difference of 10% between the treatments is such a difference truly exists, (11.3.10) can be used to calculate the required sample size as follows:

$$\begin{aligned} n &= \frac{[Z(\alpha/2)\sqrt{2P(1-P)} + Z(\beta)\sqrt{P_1(1-P_1) + P_2(1-P_2)}]^2}{(P_1 - P_2)^2} \\ &= \frac{[(1.96)\sqrt{2(0.85)(0.15)} + 0.842\sqrt{0.9(0.1) + 0.8(0.2)}]^2}{(0.1)^2} \\ &= 199.02 \approx 200. \end{aligned}$$

Therefore a sample size of 200 patients per arm is required to obtain an 80% power for detection of a difference of 10% between treatment response rates.

11.4 MULTIPLE SAMPLES

Sample Size Calculations for Analysis of Variance Models

For one-way analysis of variance with n observations per treatment, the main objective is to test

$$\begin{aligned} H_0: \tau_1 &= \tau_2 = \dots = \tau_k, \\ \text{vs. } H_a: &\text{At least one of } \tau_i \text{'s are not zero.} \end{aligned}$$

Recall that

$$\begin{aligned} E(\text{MSA}) &= E\left(\frac{\text{SSA}}{k-1}\right) = \sigma^2 + \frac{n}{k-1} \sum_{i=1}^k \tau_i^2, \\ E(\text{MSE}) &= E\left(\frac{\text{SSE}}{k(n-1)}\right) = \sigma^2. \end{aligned}$$

Thus, for a given deviation from the null hypothesis H_0 , as measured by $n \sum_{i=1}^k \tau_i^2 / (k-1)$, large value of σ^2 decreases the chance of obtaining a value $F_A = \text{MSA}/\text{MSE}$ that is in the critical region for the test. The sensitivity of the test describes the ability of the procedure to detect differences in the population means, which is measured by the power of the test. The power is interpreted as the probability that the F statistic is in the critical region when the null hypothesis is false and the treatment means differ. Since under the null hypothesis, $F_A = \text{MSA}/\text{MSE}$ follows an F distribution with (v_1, v_2) degrees of freedom,

where $v_1 = k - 1$ and $v_2 = k(n - 1) = N - k$. For the one-way analysis of variance, the power is given by

$$\begin{aligned} 1 - \beta &= P\{F_A > f(\alpha, v_1, v_2) | H_a \text{ is true}\} \\ &= P\left\{F_A > f(\alpha, v_1, v_2) \mid \frac{n}{k-1} \sum_{i=1}^k \tau_i^2\right\}. \end{aligned} \quad (11.4.1)$$

For given values of $n \sum_{i=1}^k \tau_i^2 / (k - 1)$ and σ^2 , the power can be increased by using a large sample size. The problem becomes one of designing the experiment with a value of n so that the power requirements are met. Under the alternative hypothesis that $\sum_{i=1}^k \tau_i^2 \neq 0$, F_A follows a noncentral F distribution with a noncentrality parameter δ where

$$\delta^2 = \frac{n \sum_{i=1}^k \tau_i^2}{2\sigma^2}.$$

Thus (11.4.1) becomes

$$\begin{aligned} 1 - \beta &= P\left\{F_A > f(\alpha, v_1, v_2) \mid \frac{n}{k-1} \sum_{i=1}^k \tau_i^2\right\} \\ &= P\{F_A > f(\alpha, v_1, v_2, \delta)\}. \end{aligned} \quad (11.4.2)$$

As a result the required sample size per group can be determined because v_2 and δ are functions of n . As it can be seen from the above expression, there exists no explicit form for the required sample size n . As alternative, we can consider a normal approximation proposed by Laubscher (1960) for solution of n . The idea is to consider the following approximation:

$$Z = \frac{\sqrt{\frac{v_1(2v_2-1)F_A}{v_2}} - \sqrt{2(v_1 + \delta^2)} - \frac{v_1 + 2\delta^2}{v_1 + \delta^2}}{\sqrt{\frac{v_1 F_A}{v_2} + \frac{v_1 + 2\delta^2}{v_1 + \delta^2}}} \quad (11.4.3)$$

which is approximately normally distributed with mean zero and standard deviation one. From (11.4.3) the required sample size can be determined by solving the following equation:

$$z(\beta) = \frac{\sqrt{v_2[2(v_1 + \delta^2)^2 - (v_1 + 2\delta^2)]} - \sqrt{v_1(v_1 + \delta^2)(2v_2 - 1)F_A^*}}{\sqrt{v_1(v_1 + \delta^2)F_A^* + v_2(v_1 + 2\delta^2)}}, \quad (11.4.4)$$

where $F_A^* = F(\alpha, v_1, v_2)$.

Note that tables for the solution of (11.4.2) have been constructed by many researchers such as Kastenbaum, Hoel, and Bowman (1970) and Cohen (1977). Charts for solutions of (11.4.2) have also been developed by many authors such as Pearson and Hartley (1951) and Feldt and Mahmoud (1958). For example, consider the use of the charts developed by Pearson and Hartley (1951). For convenience sake, define

$$\lambda = \frac{n \sum_{i=1}^k \tau_i^2}{2\sigma^2},$$

and let $\phi^2 = 2\lambda/(v_1 + 1)$. Then λ and ϕ^2 can be expressed as

$$\lambda = \frac{v_1 E(\text{MSA})}{2\sigma^2} - \frac{v_1}{2},$$

$$\phi^2 = \frac{v_1}{v_1 + 1} \frac{E(\text{MSA}) - \sigma^2}{\sigma^2}.$$

Table 11.4.1 gives the power of the analysis of variances as a function of ϕ for various values of v_1 , v_2 , and the significance level α . This table can be used to determine the sample size for other fixed effects models such as the randomized complete block model in which λ and ϕ are given. For example, for the randomized complete block model,

$$\lambda = \frac{b \sum_{i=1}^K \tau_i^2}{2\sigma^2},$$

$$\phi^2 = \frac{b \sum_{i=1}^K \tau_i^2}{k\sigma^2}.$$

Example 11.4.1 To illustrate the use of (11.4.4) for the sample size determination in comparing more than two treatments, consider the following example: Suppose that we are interested in conducting a four-arm parallel group, double-blind, randomized clinical trial to compare four treatments. The comparison will be made based on an F test with a significance level of $\alpha = 0.05$. Assume that the standard error within each group is expected to be $\sigma = 3.5$ and that the clinically important differences for the four treatment groups are given by

$$\tau_1 = -0.75, \quad \tau_2 = 3.0, \quad \tau_3 = -0.5, \quad \text{and} \quad \tau_4 = -1.75$$

Thus we have

$$\begin{aligned} \delta^2 &= \frac{n \sum_{i=1}^K \tau_i^2}{\sigma^2} \\ &= \frac{n[(0.75)^2 + (3.0)^2 + (0.5)^2 + (1.75)^2]}{(3.5)^2} \\ &= 1.092n. \end{aligned}$$

The sample size can be determined using (11.4.4). To obtain the required sample size, we apply various n to (11.4.4). The results are summarized in Table 11.4.2. From Table 11.4.2, it can be seen that for $n = 11$ and $F^* = F(0.05, 3, 40) = 2.8387$, equation (11.4.4) yields $z(\beta) \approx 0.853$ which is the closest to $z(0.2) = 0.842$. Therefore $n = 11$ is the required sample size per treatment group. Note that the required sample size $n = 11$ can also be obtained from Table 11.4.1 by specifying λ and ϕ .

Sample Size Calculations for Generalized Linear Models

In many cases test statistics for the null hypothesis of no treatment differences can be derived within the framework of generalized linear models. Self and Mauritsen (1988) propose an approach for sample size and power calculations within the framework of generalized linear

Table 11.4.1 Power of the Analysis of Variance Test

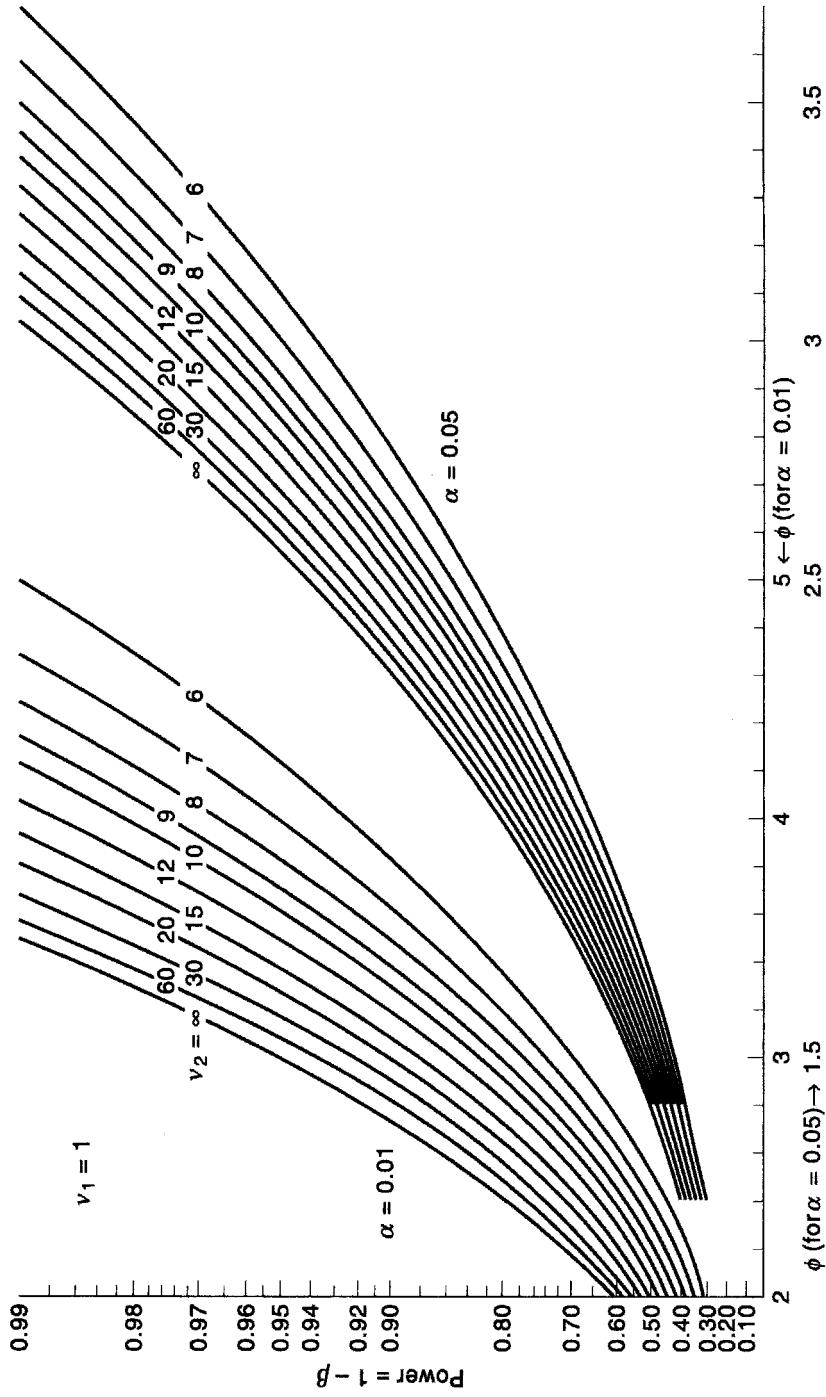


Table 11.4.1 (Continued)

456

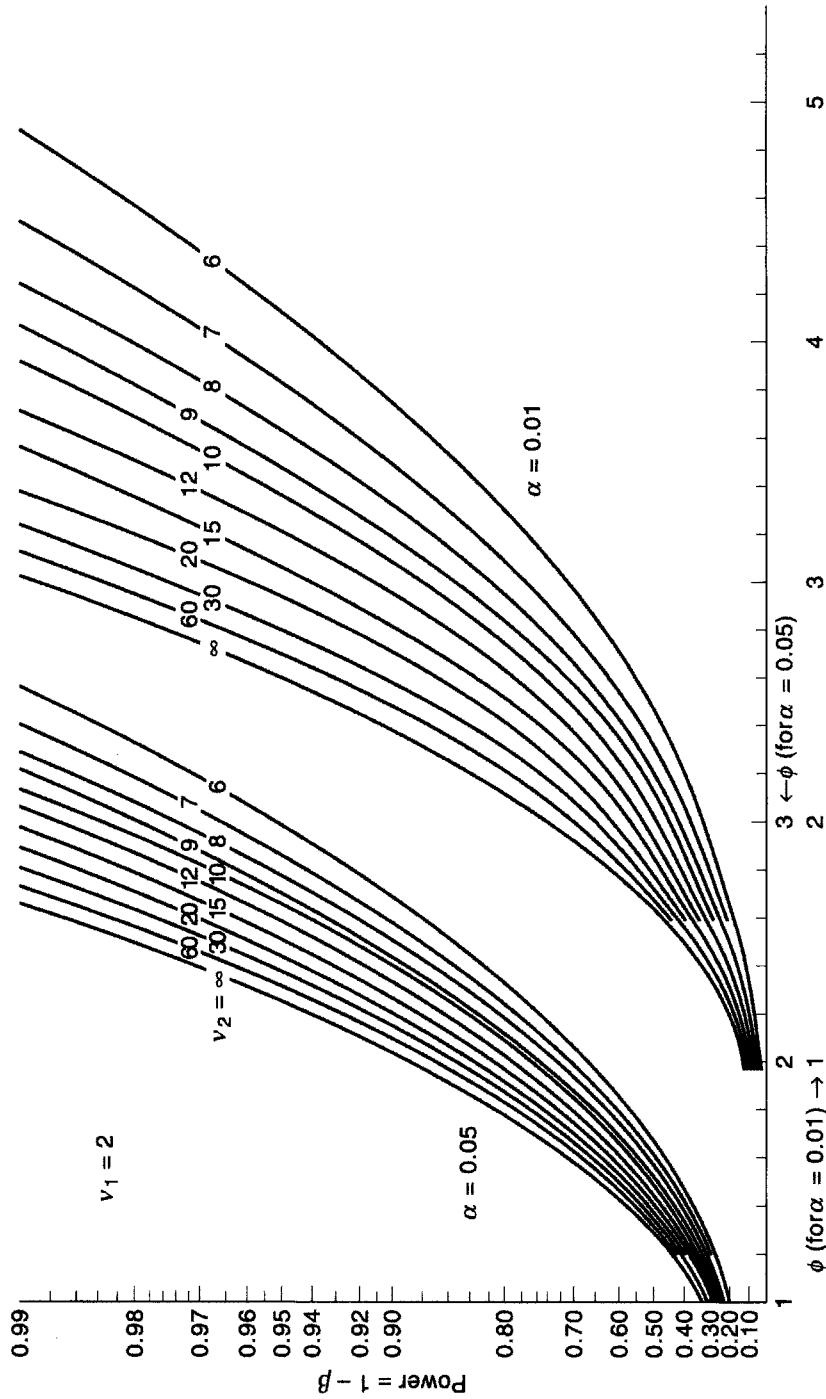


Table 11.4.1 (Continued)

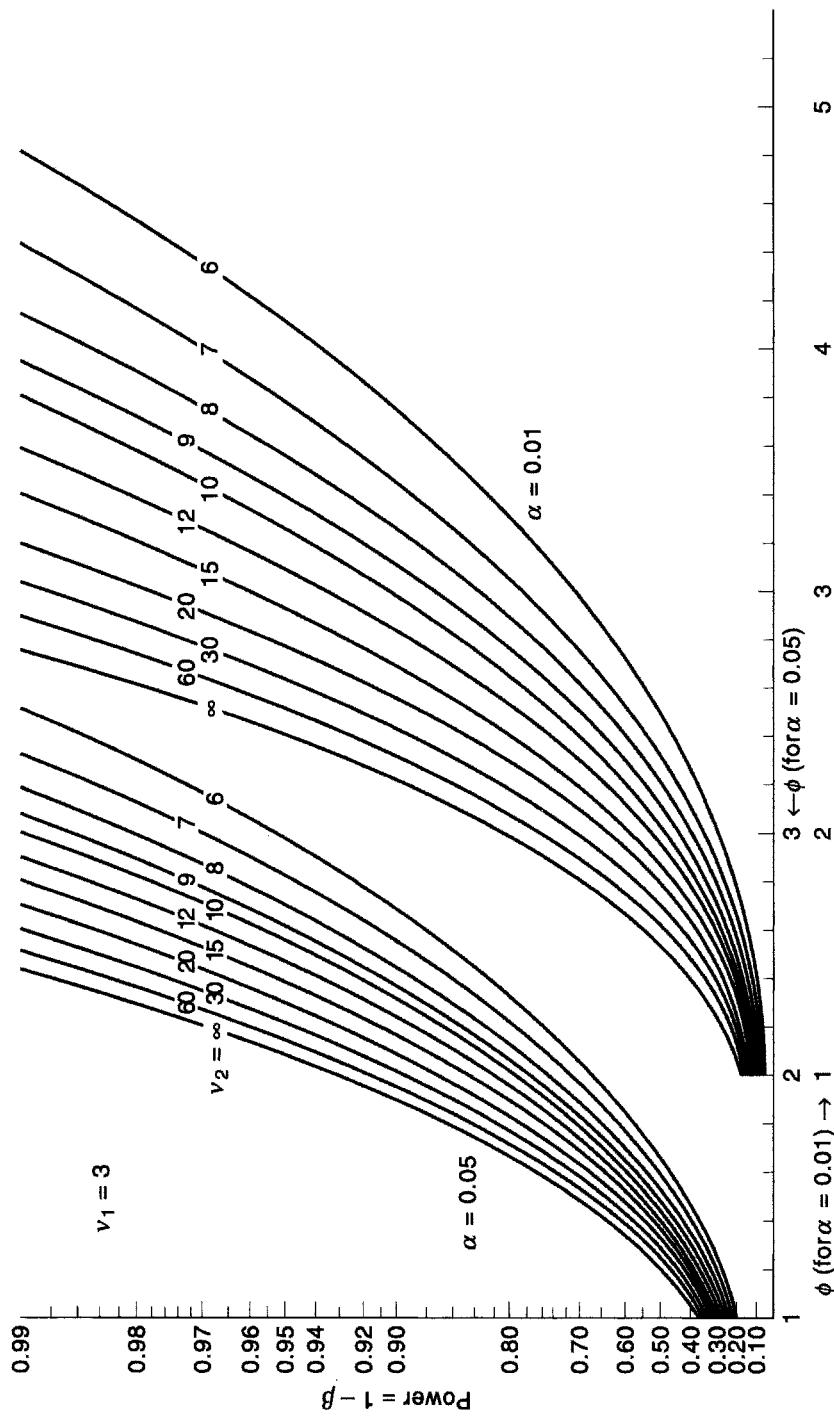


Table 11.4.1 (Continued)

458

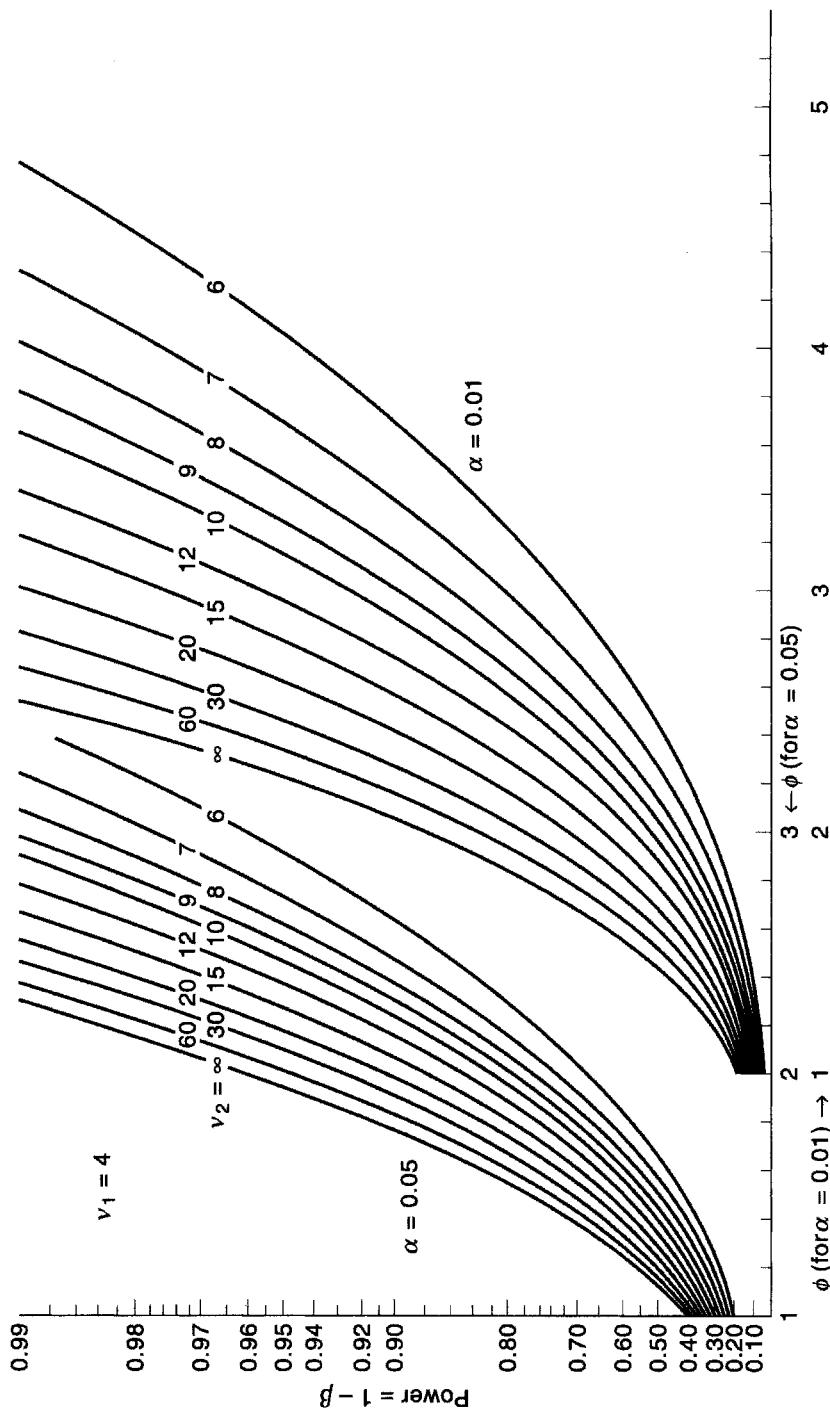


Table 11.4.1 (Continued)

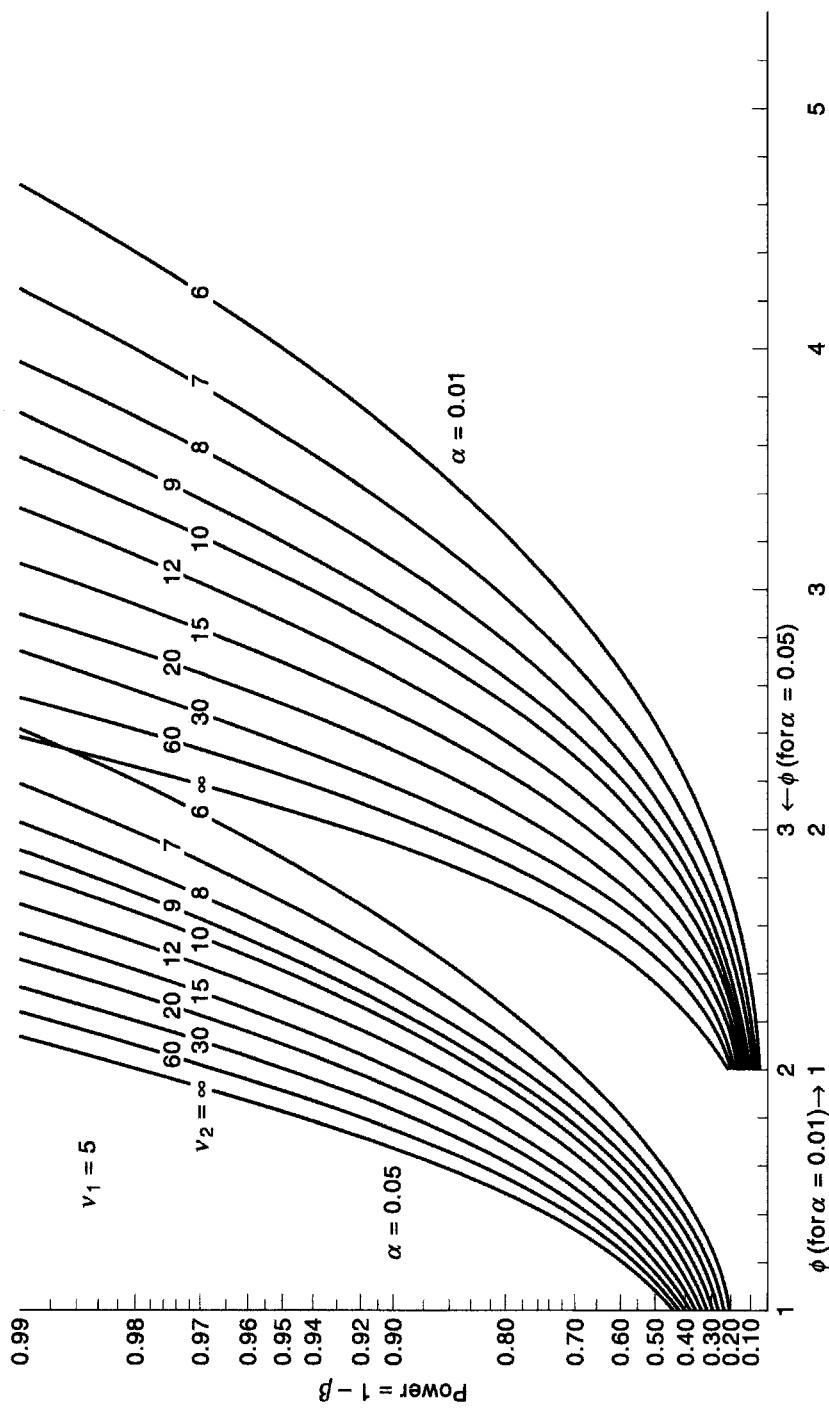


Table 11.4.1 (Continued)

460

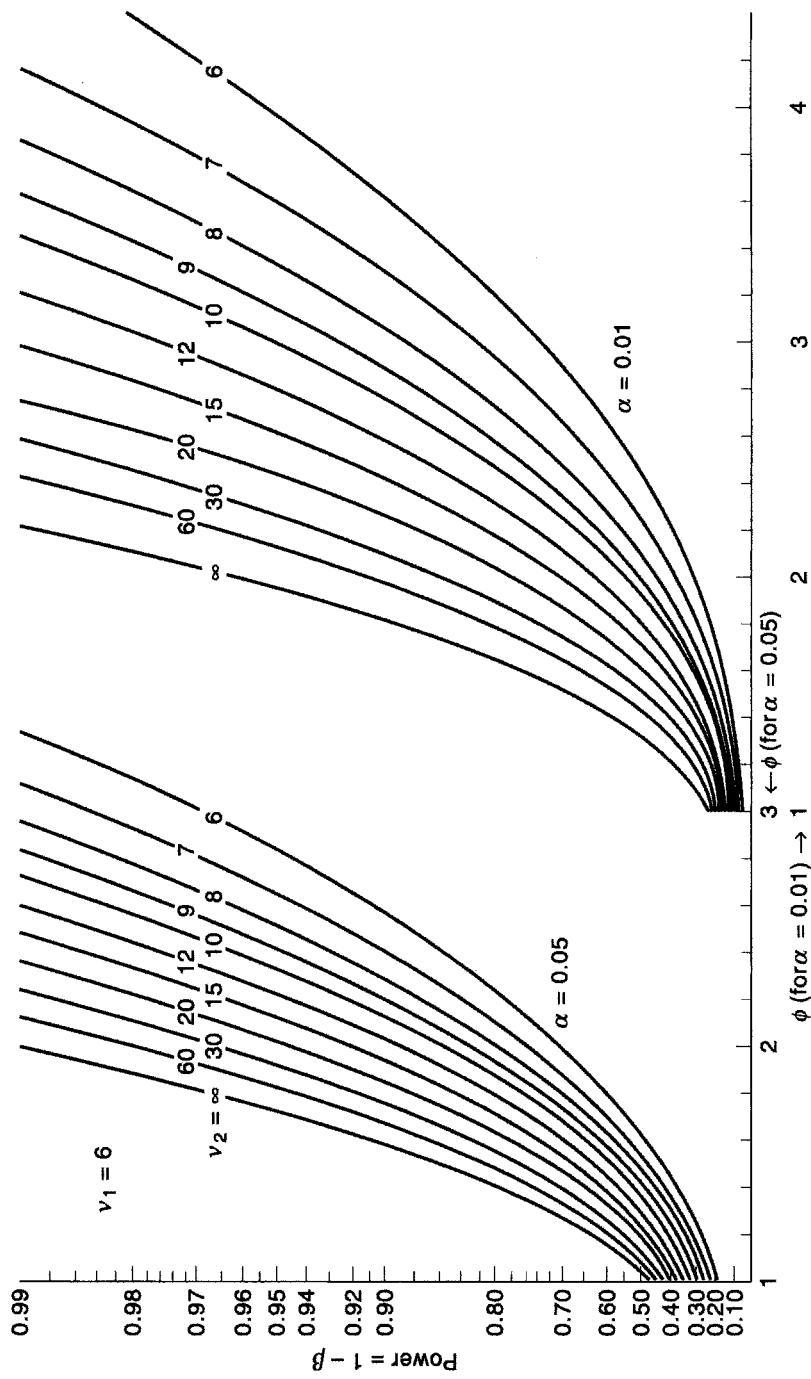


Table 11.4.1 (Continued)

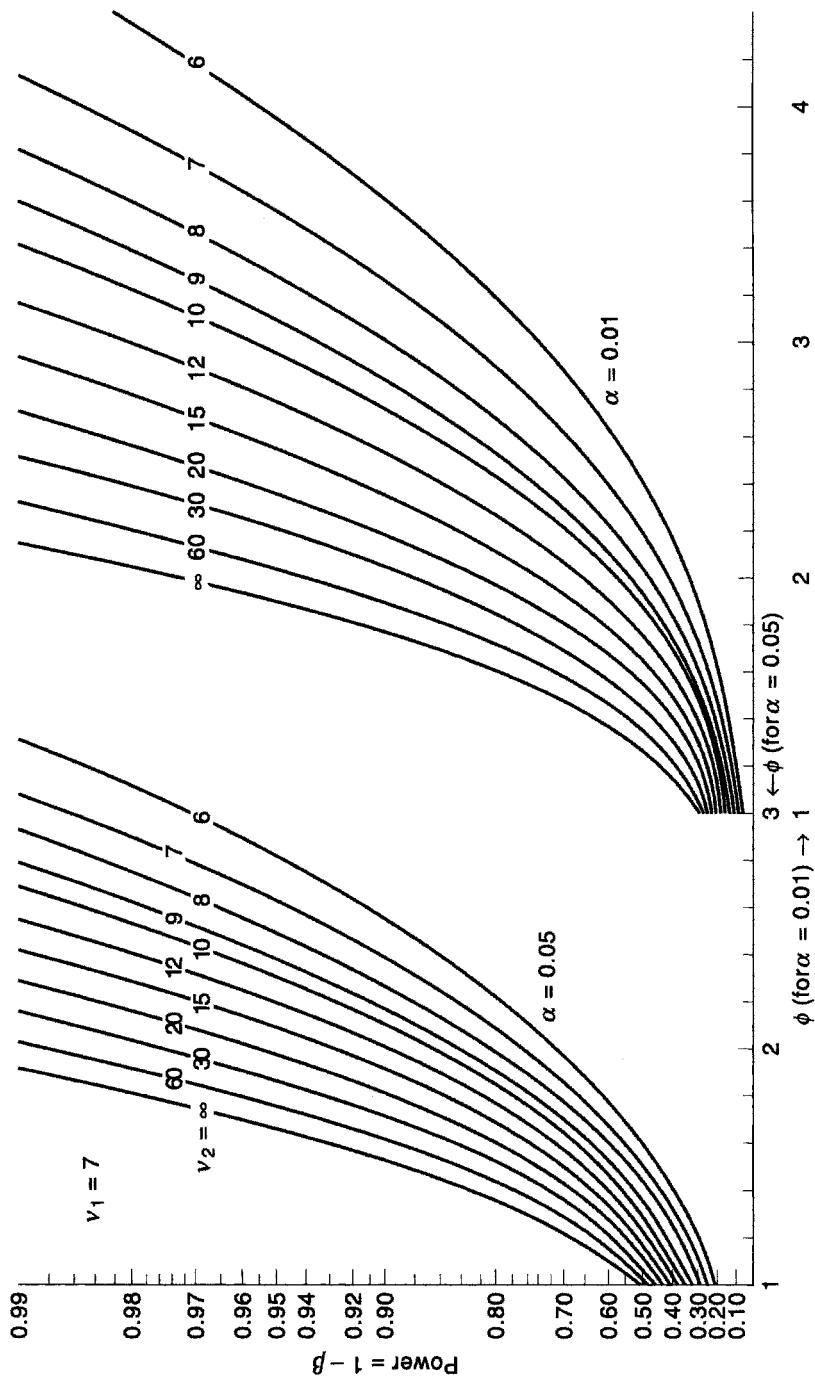
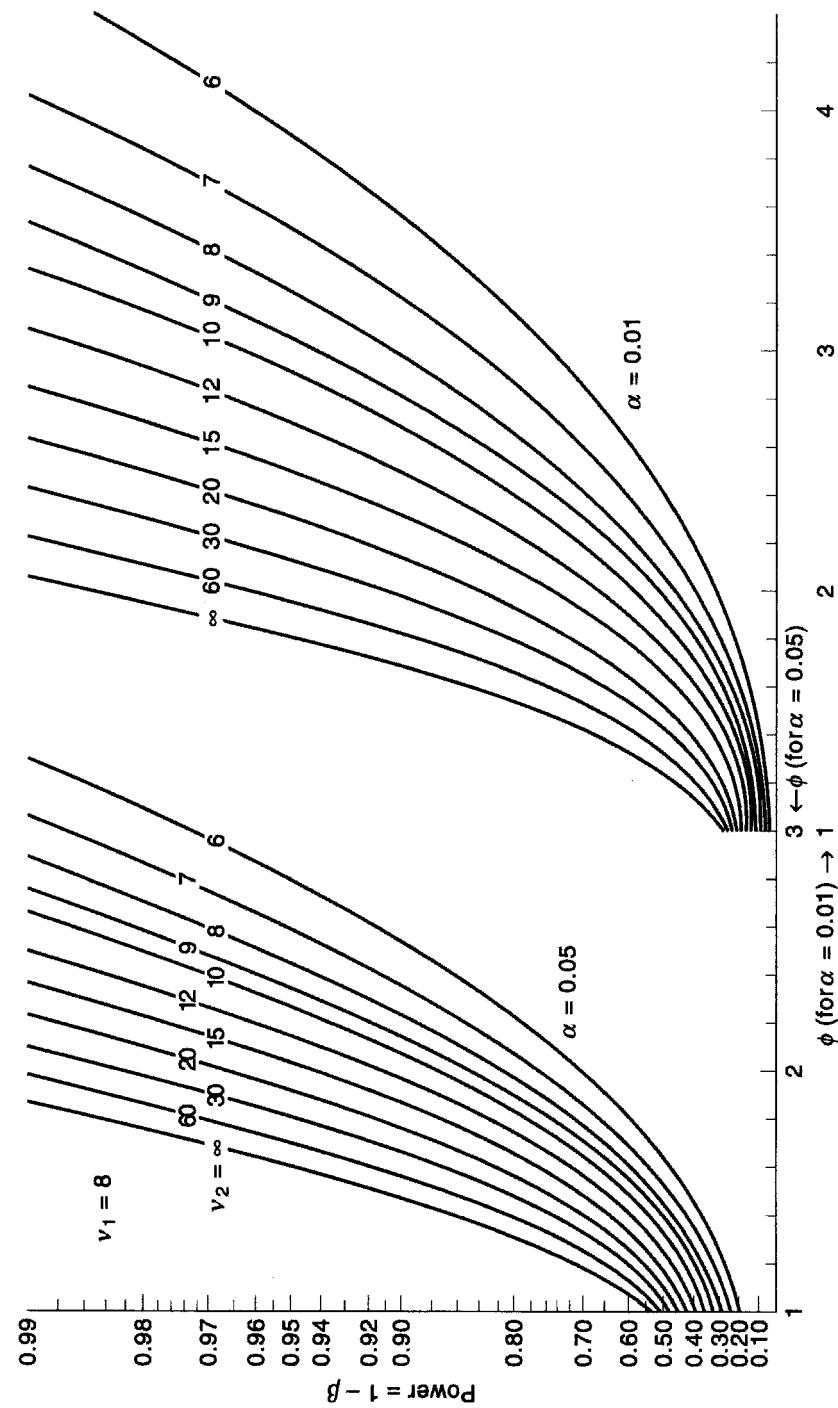


Table 11.4.1 (Continued)



Source: Reproduced from E. S. Pearson and H. O. Hartley, Charts of the power function for analysis-of-variance tests, derived from the noncentral F distribution, *Biometrika*, **38**, 1951, by permission of the editor.

Table 11.4.2 Sample Size Calculation for Multiple Samples

<i>n</i>	<i>v</i> ₁	<i>v</i> ₂	<i>F</i> _A [*]	<i>Z</i> (β)
9	3	32	2.901	0.524
10	3	36	2.866	0.693
11	3	40	2.839	0.853
12	3	44	2.817	1.006
13	3	48	2.798	1.152
14	3	52	2.783	1.293

Note: $Z(\beta)$ are obtained based on(11.4.4).

models (e.g., McCullagh and Nelder, 1989) that is based on a noncentral chi-square approximation to the distribution of score statistics. In addition, Self, Mauritsen, and Ohara (1992) derive a more accurate approach that is based on a noncentral chi-square approximation to the distribution of the likelihood ratio statistic. Their approaches provide a unified tool for sample size calculations for continuous and discrete responses. In general, there exist no explicit formulas to use in determining the sample size for a generalized linear model. Hence the numerical methods are usually required. These methods have been implemented in EGRET SIZ (SERC, 1993). The SIZ of EGRET not only calculates sample size for a given model but also performs a simulation to compare the empirical power with respect to the nominal one. This provides a way to assess the performance of the calculated sample size in practice.

Liu and Liang (1995) propose a method to compute sample size and power with correlated observations. Let Y_{ij} be the j th repeated measurement of a response variable for the i th subject, where $i = 1, \dots, n$ and $j = 1, \dots, m$. Then Y_{ij} can be described by the following linear regression model:

$$Y_{ij} = X_{ij}\boldsymbol{\psi} + Z_{ij}\boldsymbol{\lambda} + \varepsilon_{ij},$$

where X_{ij} and Z_{ij} represent the design matrix and covariates, respectively, $\boldsymbol{\psi}$ is a $p \times 1$ vector of parameters of interest, and $\boldsymbol{\lambda}$ is a $q \times 1$ vector of nuisance parameters. It is assumed that $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{im})'$ follows a multivariate normal with mean $\mathbf{0}$ and covariance matrix $\sigma^2\boldsymbol{\Sigma}$. Under the above model, if we consider the special case where $p = q = 1$ with $Z_{ij} = 1$ for all i and j and $X_{ij} = X_i = 1$ (treatment) or 0 (placebo), then it reduces to a typical two samples problem with repeated measures. In this case, the sample size required can be obtained as

$$n = \frac{[Z(\alpha/2) + Z(\beta)]^2 \sigma^2 [1 + (m - 1)\rho]}{\pi(1 - \pi)\psi_1^2 m}, \quad (11.4.5)$$

where ρ is the within-subject correlation, π is the proportion of subjects in the treatment group, and $\psi_1 = \Delta$ is the difference of clinical importance. The above formula can be applied to the case where two treatments are compared with binary responses with the following modification:

$$n = \frac{[Z(\alpha/2) + Z(\beta)]^2 [\pi P_0(1 - P_0) + (1 - \pi)P_1(1 - P_1)][1 + (m - 1)\rho]}{m\pi(1 - \pi)(P_1 - P_0)^2}, \quad (11.4.6)$$

where P_0 and P_1 are response probabilities for the placebo and the treatment groups, respectively.

11.5 CENSORED DATA

For survival analysis in clinical trials, in order to compare the median survival time to some event between two populations, some assumptions are essential (Lachin, 1981). Let $P_i(t)$ denote the probability of surviving to time t for subjects in the i th population, and assume exponential survival with hazard rate λ_i , $i = 1, 2$. Then

$$P_i(t) = e^{-\lambda_i t},$$

and the median survival time is

$$M_i = \frac{\ln 2}{\lambda_i}.$$

As indicated by Lachin (1981), if we further assume uniform censoring time T_0 from the start of the study to termination of follow-up at time T after the start of the study, then

$$\hat{\lambda}_i = \frac{\text{Number of events in sample for population } i}{\text{Total follow-up time for population } i}$$

approximately follows a normal distribution with mean λ_i and variance $\phi(\lambda_i)/n$, where

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

Hence the required sample size for testing $H_0: M_1 = M_2$ versus $H_a: M_2 > M_1$ using a normal approximation is given by

$$n = \frac{\left[z(\alpha)\sqrt{2\phi(\bar{\lambda})} + z(\beta)\sqrt{\phi(\lambda_1) + \phi(\lambda_2)} \right]^2}{(\lambda_2 - \lambda_1)^2}. \quad (11.5.1)$$

Sample size determination in clinical trials with censored endpoints is complicated by the fact that the risk of event for patients may not remain constant during the trial. The comparison of survival curves between treatments is often performed based on censored data to determine whether a test treatment can reduce mortality rate or improve survival probability as compared to a placebo or a standard therapy. Let F and G be the failure time distributions in the treatment and placebo groups. Then, the hypotheses of interest as given in (10.3.1) are

$$\begin{aligned} H_0: 1 - F &= 1 - G, \\ \text{vs. } H_a: 1 - F &\neq 1 - G. \end{aligned} \quad (11.5.2)$$

In a similar manner the sample size can be determined based on a power analysis of a test statistic for the above hypotheses. For testing (11.5.2), several tests have been proposed in literature. For example, George and Desu (1973) and Rubinstein, Gail, and Santner (1981) discuss tests based on exponential survival curves. Tests based on exponential survival curves, however, are derived under the very restrictive assumption of constant hazard

ratios. In practice, the hazard rate, which is often time dependent, can vary even if the effect of therapy is constant over time. As a result the proportional hazards assumption is not valid (Wu, Fisher, and DeMets, 1980; Lachin and Foulkes, 1986). Therefore, the usual tests based on exponential models with constant hazard ratios no longer apply.

As an alternative, the logrank test is often considered. Schoenfeld (1981) and Freedman (1982) present methods for sample size calculation based on the asymptotic expectation and variance of the logrank statistic. The conditions under which the formulas for sample size are derived, however, are very restrictive. In this section, we will introduce a sample size formula by Lakatos (1986, 1988) which is derived under very general conditions. As was discussed in Section 10.3, the logrank test statistic is a special case of a general class of nonparametric tests for comparing two survival functions. Lakatos (1988) considered expressing the logrank statistic as a member of the Tarone-Ware class of statistics as follows:

$$T = \frac{\sum_{k=1}^d w_k \left(X_k - \frac{n_{2k}}{n_{2k} + n_{1k}} \right)}{\sum_{k=1}^d w_k^2 \left[\frac{n_{2k} n_{1k}}{(n_{2k} + n_{1k})^2} \right]^{1/2}}, \quad (11.5.3)$$

where $d = \sum_{k=1}^K d_k$ is the total of deaths, X_k is the indicator of the placebo, w_k is the k th Tarone-Ware weight, and n_{jk} are the numbers at risk just before the k th death in the j th treatment. Under a fixed local alternative, Lakatos (1988) indicates that the expectation of (11.5.3) can be approximated by

$$E = \frac{\sum_{i=1}^N \sum_{k=1}^{d_i} w_{ik} \left[\frac{\phi_{ik} \theta_{ik}}{1 + \phi_{ik} \theta_{ik}} - \frac{\phi_{ik}}{1 + \phi_{ik}} \right]}{\left[\sum_{i=1}^N \sum_{k=1}^{d_i} \frac{w_{ik}^2 \phi_{ik}}{(1 + \phi_{ik})^2} \right]^{1/2}}, \quad (11.5.4)$$

where $N = y(K)$, ϕ_{ik} is the ratio of patients in the two treatment groups at risk just prior to the k th death in the $y(i)$ th interval,

$$\theta_{ik} = \frac{P_{1ik}}{P_{2ik}},$$

where P_{jik} is the hazard just prior to the k th death in $y(i)$ th interval in the j th treatment, and w_{ik} is the corresponding Tarone-Ware weight. When $w_{ik} = 1$ for all i and k , (11.5.3) reduces to the logrank test. Treating this statistic as a normal random variable with mean E and variance 1, we have

$$E = Z(\alpha/2) + Z(\beta).$$

Assuming that $\phi_{ik} = \phi_i$ and $w_{ik} = w_i$ for all k in the $y(i)$ th interval and letting $\rho_i = d_i/d$, then (11.5.4) becomes

$$E = \frac{\sqrt{d} \sum_{i=1}^N w_i \rho_i \gamma_i}{\left(\sum_{i=1}^N w_i^2 \rho_i \eta_i \right)^{1/2}},$$

where

$$\gamma_i = \frac{\phi_i \theta_i}{1 + \phi_i \theta_i} - \frac{\phi_i}{1 + \phi_i}$$

and

$$\eta_i = \frac{\phi_i}{(1 + \phi_i)^2}.$$

This leads to

$$d = \frac{[Z(\alpha/2) + Z(\beta)]^2 \sum_{i=1}^N w_i^2 \rho_i \eta_i}{\left(\sum_{i=1}^N w_i \rho_i \gamma_i \right)^2}. \quad (11.5.5)$$

Since

$$d = \frac{n(P_1 + P_2)}{2},$$

where P_1 and P_2 are cumulative event rates for treatment and placebo groups, respectively, the required total sample size can be obtained as

$$n = \frac{2d}{P_1 + P_2}. \quad (11.5.6)$$

Yateman and Skene (1992) also propose a method for sample size determination for proportional hazards survival studies with arbitrary patient entry and loss to follow-up distributions.

Example 11.5.1 To illustrate the above methodology for sample size determination, consider the example described in Gail (1985). Suppose that we have a two-year trial with event rates of $1 - \exp(-1) \approx 0.6321$ and $1 - \exp(-\frac{1}{2}) \approx 0.3935$ per year in the placebo and treatment groups, respectively, and that the yearly loss to follow-up and noncompliance rates are 3% and 4%, respectively. Assuming that hazard rate is constant. The quantities ρ_i , η_i , and γ_i can readily be determined using the Markov model as described in Lakatos (1988) which accounts for lost to follow-up, noncompliance, lag time, and so forth. Table 11.5.1 displays these parameters including ϕ_i and θ_i . By (11.5.4) it can be verified that the number of deaths d is 102. In addition, from the last row of Table 11.5.1, the cumulative event rates over a two-year period are given by 83.7% and 62.7% for the treatment and placebo groups, respectively. Thus, the required total sample size for achieving a 90% power at 5% level of significance can be obtained as

$$n = \frac{2(102)}{0.837 + 0.627} = 139.$$

Note that a SAS computer program implementing the Markov models and some variations is given by Lakatos (1986). In addition, Shih (1995) illustrate the use of SIZE, a comprehensive computer program for calculating sample size, power, and duration of study in clinical trials with time-dependent rates of event, crossover, and loss to follow-up. As indicated by Shih (1995), SIZE covers a wide range of complexities commonly occurring in clinical trials such as nonproportional hazards, lag in treatment effect, and uncertainties in treatment benefit.

Table 11.5.1 Parameters for Example 11.5.1

t_i	Control			Experimental									
	L	E	A_C	A_E	L	E	A_E	A_C	γ	η	ρ	θ	ϕ
0.1	0.003	0.095	0.897	0.005	0.003	0.049	0.944	0.004	0.167	0.222	0.098	2.000	1.000
0.2	0.006	0.181	0.804	0.009	0.006	0.095	0.891	0.007	0.166	0.226	0.090	1.986	0.951
0.3	0.008	0.258	0.721	0.013	0.009	0.139	0.842	0.010	0.166	0.230	0.083	1.972	0.905
0.4	0.010	0.327	0.647	0.016	0.011	0.181	0.795	0.013	0.165	0.234	0.076	1.959	0.862
0.5	0.013	0.389	0.580	0.018	0.014	0.221	0.750	0.015	0.164	0.237	0.070	1.945	0.821
0.6	0.014	0.445	0.520	0.020	0.016	0.259	0.708	0.016	0.163	0.240	0.064	1.932	0.782
0.7	0.016	0.496	0.466	0.022	0.018	0.295	0.669	0.017	0.162	0.242	0.059	1.920	0.746
0.8	0.017	0.541	0.418	0.023	0.020	0.330	0.632	0.018	0.160	0.244	0.054	1.907	0.711
0.9	0.019	0.582	0.375	0.024	0.022	0.362	0.596	0.019	0.158	0.246	0.050	1.894	0.679
1.0	0.020	0.619	0.336	0.024	0.024	0.393	0.563	0.019	0.156	0.248	0.046	1.882	0.648
1.1	0.021	0.652	0.302	0.025	0.026	0.423	0.532	0.020	0.154	0.249	0.043	1.870	0.619
1.2	0.022	0.682	0.271	0.025	0.028	0.450	0.502	0.020	0.152	0.249	0.039	1.857	0.592
1.3	0.023	0.709	0.243	0.025	0.029	0.477	0.474	0.020	0.149	0.250	0.036	1.845	0.566
1.4	0.024	0.734	0.218	0.025	0.031	0.502	0.448	0.020	0.147	0.250	0.034	1.833	0.542
1.5	0.025	0.755	0.195	0.025	0.032	0.525	0.423	0.020	0.144	0.250	0.031	1.820	0.519
1.6	0.025	0.775	0.175	0.024	0.033	0.548	0.399	0.019	0.141	0.249	0.029	1.808	0.497
1.7	0.026	0.793	0.157	0.024	0.035	0.569	0.377	0.019	0.138	0.248	0.027	1.796	0.477
1.8	0.026	0.809	0.141	0.023	0.036	0.589	0.356	0.018	0.135	0.247	0.025	1.783	0.457
1.9	0.027	0.824	0.127	0.023	0.037	0.609	0.336	0.018	0.132	0.246	0.023	1.771	0.439
2.0	0.027	0.837	0.114	0.022	0.038	0.627	0.318	0.018	0.129	0.244	0.021	1.758	0.421

11.6 DOSE-RESPONSE STUDIES

As indicated by Ruberg (1995a, 1995b), the following fundamental questions that dictate design and analysis strategies are necessarily addressed when studying the dose-response relationship of a new drug:

1. Is there any drug effect?
2. What doses exhibit a response different from control?
3. What is the nature of the dose-response relationship?
4. What is the optimal dose?

The first two questions are usually addressed by means of the techniques of analysis of variance, and the last two questions are related to the identification of minimum effective dose (MED). In this section, we will introduce sample size calculations based on the concept of the dose-response study using analysis of variance and statistical test for MED.

Dose-Response Relationship

In a dose-response study suppose that there is a control group and K dose groups. The null hypothesis of interest is then given by

$$H_0: \mu_0 = \mu_1 = \dots = \mu_K, \quad (11.6.1)$$

where μ_0 is mean response for the control group and μ_i is mean response for the i th dose group. The rejection of hypothesis (11.6.1) indicates that there is a drug effect. The dose-response relationship can then be examined under appropriate alternative hypotheses. Under a specific alternative hypothesis, the required sample size per dose group can then be obtained. Spriet and Dupin-Spriet (1996) identify the following alternative hypotheses for dose response:

1. $H_a: \mu_0 < \mu_1 < \dots < \mu_{K-1} < \mu_K$.
2. $H_a: \mu_0 < \dots < \mu_i = \dots = \mu_j > \dots > \mu_K$.
3. $H_a: \mu_0 < \dots < \mu_i = \dots = \mu_K$.
4. $H_a: \mu_0 = \mu_1 < \dots < \mu_K$.
5. $H_a: \mu_0 < \mu_1 < \dots = \mu_i = \dots = \mu_K$.
6. $H_a: \mu_0 = \dots = \mu_i < \dots < \mu_{K-1} < \mu_K$.
7. $H_a: \mu_0 = \mu_1 < \dots < \mu_i = \dots = \mu_K$.
8. $H_a: \mu_0 = \dots = \mu_i < \dots < \mu_{K-1} = \mu_K$.

Figure 11.6.1 exhibits dose-response patterns under the above alternative hypotheses. Under these alternative hypotheses the statistical tests can be very complicated, and hence there may exist no close form for the corresponding power functions. As an alternative, Spriet and Dupin-Spriet (1996) employ simulations to obtain adequate sample sizes for parallel-group dose-response clinical trials. An example of an alternative hypothesis on linear contrast is

$$H_a: \sum_{j=0}^K c_j \mu_j \neq 0,$$

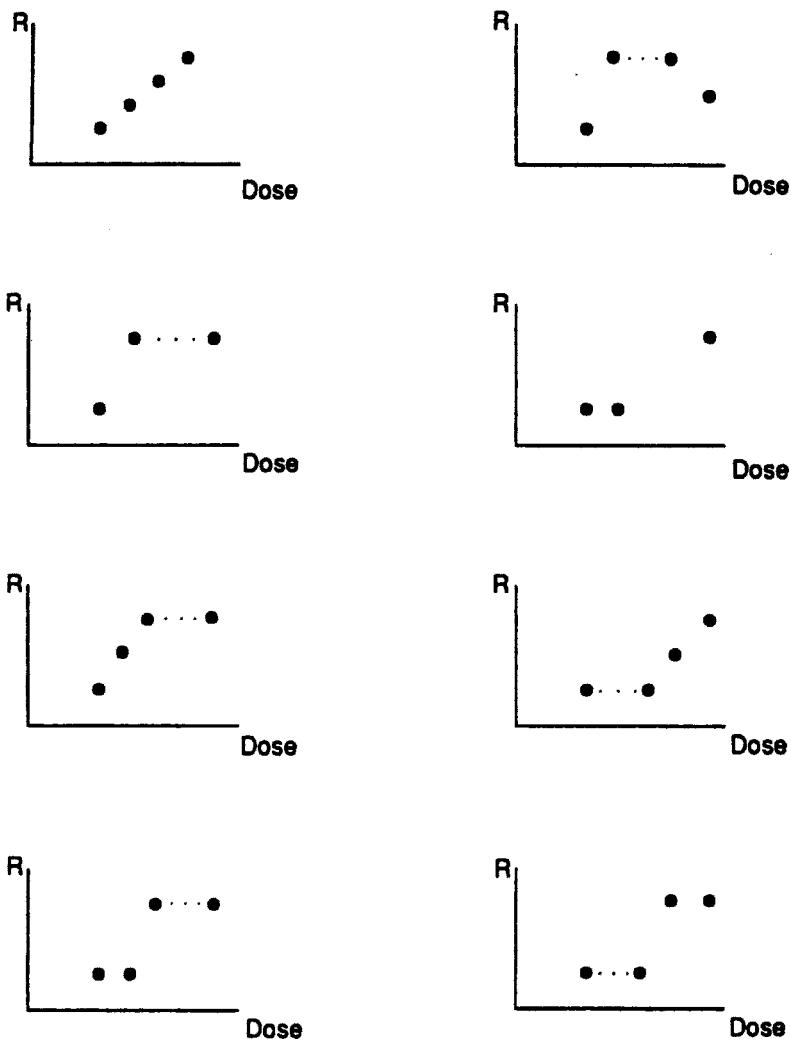


Figure 11.6.1 Commonly encountered dose-response patterns.

where

$$\sum_{j=0}^K c_j = 0.$$

The given contrast of means is

$$C = \sum_{j=0}^K c_j \mu_j;$$

it can be estimated by the contrast of sample means

$$\hat{C} = \sum_{j=0}^K c_j \bar{Y}_j,$$

where \bar{Y}_j is the mean response of n patients receiving the j th dose level. Thus the statistic

$$T = \frac{\sum_{j=0}^K c_j \bar{Y}_j}{\sqrt{\hat{\sigma}^2 \sum_{j=0}^K c_j^2 / n}},$$

can be used to test the significance of the dose response of contrast C , where $\hat{\sigma}^2$ is the mean square error from the analysis of variance model used to estimate C . Under the alternative hypothesis, the contrast power is given by

$$\begin{aligned} 1 - \beta &= \Phi \left[Z(\alpha/2) - \frac{\sum_{j=0}^K c_j \mu_j}{\sqrt{\sigma^2 \sum_{j=0}^K c_j^2 / n}} \right] \\ &\quad + 1 - \Phi \left[Z(1 - \alpha/2) - \frac{\sum_{j=0}^K c_j \mu_j}{\sqrt{\sigma^2 \sum_{j=0}^K c_j^2 / n}} \right], \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative standard normal distribution function. Hence the per-dose sample size necessary for the $(1-\beta)100\%$ power to show a statistical significance in the dose response with the contrast is

$$n = \frac{[Z(\alpha/2) + Z(\beta)]^2 \hat{\sigma}^2 \sum_{j=0}^K c_j^2}{[\sum_{j=0}^K c_j \mu_j]^2} \quad (11.6.2)$$

Minimum Effective Dose

For the comparison of the mean dose treatment with the control mean, Williams (1971, 1972) proposes a test that determines the lowest dose level at which there is evidence for a difference from control. Williams considers the alternative hypothesis

$$H_a: \mu_0 = \mu_1 = \dots = \mu_{i-1} < \mu_i \leq \mu_{i+1} \leq \dots \leq \mu_K,$$

and proposes the statistic

$$T_i = \frac{\hat{\mu}_i - \bar{Y}_0}{s \sqrt{(1/n_i) + (1/n_0)}},$$

where s^2 is an unbiased estimate of σ^2 that is independent of \bar{Y}_i and is distributed as $\sigma^2 \chi^2_v / v$ and $\hat{\mu}_i$ is the maximum likelihood estimate of μ_i , which is given by

$$\hat{\mu}_i = \max_{1 \leq u \leq i} \min_{i \leq v \leq K} \left\{ \frac{\sum_{j=u}^v n_j \bar{Y}_j}{\sum_{j=1}^v n_j} \right\}.$$

When $n_i = n$ for $i = 0, 1, \dots, K$, the test statistic can be simplified as

$$T_i = \frac{\hat{\mu}_i - \bar{Y}_0}{s \sqrt{2/n}},$$

which can be approximated by

$$T_i \sim \frac{X_i - Z_0}{s},$$

where

$$X_i = \max_{1 \leq u \leq i} \sum_{j=u}^i \frac{Z_j}{i-u+1}$$

and Z_j follows a standard normal distribution. We then reject the null hypothesis (11.6.1) and conclude that dose i is the minimum effective dose if

$$T_j > t_j(\alpha) \quad \text{for all } j \geq i,$$

where $t_j(\alpha)$ is the upper α th quantile of the distribution of T_j . Note that $t_j(\alpha)$ are given in Tables 11.6.1 through 11.6.4.

Table 11.6.1 Upper 5% Points $t_k(\alpha)$ of the Distribution T_k

df	$k = \text{Number of Dose Levels}$										
	v	1	2	3	4	5	6	7	8	9	10
5		2.02	2.14	2.19	2.21	2.22	2.23	2.24	2.24	2.25	2.25
6		1.94	2.06	2.10	2.12	2.13	2.14	2.14	2.15	2.15	2.15
7		1.89	2.00	2.04	2.06	2.07	2.08	2.09	2.09	2.09	2.09
8		1.86	1.96	2.00	2.01	2.02	2.03	2.04	2.04	2.04	2.04
9		1.83	1.93	1.96	1.98	1.99	2.00	2.00	2.01	2.01	2.01
10		1.81	1.91	1.94	1.96	1.97	1.97	1.98	1.98	1.98	1.98
11		1.80	1.89	1.92	1.94	1.94	1.95	1.95	1.96	1.96	1.96
12		1.78	1.87	1.90	1.92	1.93	1.93	1.94	1.94	1.94	1.94
13		1.77	1.86	1.89	1.90	1.91	1.92	1.92	1.93	1.93	1.93
14		1.76	1.85	1.88	1.89	1.90	1.91	1.91	1.91	1.92	1.92
15		1.75	1.84	1.87	1.88	1.89	1.90	1.90	1.90	1.90	1.91
16		1.75	1.83	1.86	1.87	1.88	1.89	1.89	1.89	1.90	1.90
17		1.74	1.82	1.85	1.87	1.87	1.88	1.88	1.89	1.89	1.89
18		1.73	1.82	1.85	1.86	1.87	1.87	1.88	1.88	1.88	1.88
19		1.73	1.81	1.84	1.85	1.86	1.87	1.87	1.87	1.87	1.88
20		1.72	1.81	1.83	1.85	1.86	1.86	1.86	1.87	1.87	1.87
22		1.72	1.80	1.83	1.84	1.85	1.85	1.85	1.86	1.86	1.86
24		1.71	1.79	1.82	1.83	1.84	1.84	1.85	1.85	1.85	1.85
26		1.71	1.79	1.81	1.82	1.83	1.84	1.84	1.84	1.84	1.85
28		1.70	1.78	1.81	1.82	1.83	1.83	1.83	1.84	1.84	1.84
30		1.70	1.78	1.80	1.81	1.82	1.83	1.83	1.83	1.83	1.83
35		1.69	1.77	1.79	1.80	1.81	1.82	1.82	1.82	1.82	1.83
40		1.68	1.76	1.79	1.80	1.80	1.81	1.81	1.81	1.82	1.82
60		1.67	1.75	1.77	1.78	1.79	1.79	1.80	1.80	1.80	1.80
120		1.66	1.73	1.75	1.77	1.77	1.78	1.78	1.78	1.78	1.78
∞		1.645	1.716	1.739	1.750	1.756	1.760	1.763	1.765	1.767	1.768

Source: Williams (1971).

Table 11.6.2 Upper 2.5% Points $t_k(\alpha)$ of the Distribution T_k

df v	k = Number of Dose Levels						
	2	3	4	5	6	8	10
5	2.699	2.743	2.766	2.779	2.788	2.799	2.806
6	2.559	2.597	2.617	2.628	2.635	2.645	2.650
7	2.466	2.501	2.518	2.528	2.535	2.543	2.548
8	2.400	2.432	2.448	2.457	2.463	2.470	2.475
9	2.351	2.381	2.395	2.404	2.410	2.416	2.421
10	2.313	2.341	2.355	2.363	2.368	2.375	2.379
11	2.283	2.310	2.323	2.330	2.335	2.342	2.345
12	2.258	2.284	2.297	2.304	2.309	2.315	2.318
13	2.238	2.263	2.275	2.282	2.286	2.292	2.295
14	2.220	2.245	2.256	2.263	2.268	2.273	2.276
15	2.205	2.229	2.241	2.247	2.252	2.257	2.260
16	2.193	2.216	2.227	2.234	2.238	2.243	2.246
17	2.181	2.204	2.215	2.222	2.226	2.231	2.234
18	2.171	2.194	2.205	2.211	2.215	2.220	2.223
19	2.163	2.185	2.195	2.202	2.205	2.210	2.213
20	2.155	2.177	2.187	2.193	2.197	2.202	2.205
22	2.141	2.163	2.173	2.179	2.183	2.187	2.190
24	2.130	2.151	2.161	2.167	2.171	2.175	2.178
26	2.121	2.142	2.151	2.157	2.161	2.165	2.168
28	2.113	2.133	2.143	2.149	2.152	2.156	2.159
30	2.106	2.126	2.136	2.141	2.145	2.149	2.151
35	2.093	2.112	2.122	2.127	2.130	2.134	2.137
40	2.083	2.102	2.111	2.116	2.119	2.123	2.126
60	2.060	2.078	2.087	2.092	2.095	2.099	2.101
120	2.037	2.055	2.063	2.068	2.071	2.074	2.076
∞	2.015	2.032	2.040	2.044	2.047	2.050	2.052

Source: Williams (1972).

Since the power function of the above test is rather complicated, as an alternative, we may consider the following approximation to obtain the required sample size per-dose group:

$$\begin{aligned}
1 - \beta &= P\{\text{reject } H_0 | \mu_i \geq \mu_0 + \Delta \text{ for some } i\} \\
&> P\{\text{reject } H_0 | \mu_0 = \mu_1 = \dots = \mu_{K-1}, \mu_K = \mu_0 + \Delta\} \\
&\geq P\left\{ \frac{\bar{Y}_K - \bar{Y}_0}{\sigma\sqrt{2/n}} > t_K(\alpha/2) | \mu_K = \mu_0 + \Delta \right\} \\
&= P\left\{ Z > t_K(\alpha/2) - \frac{\Delta}{\sigma\sqrt{2/n}} \right\},
\end{aligned}$$

where Δ is the clinically meaningful difference. If a one-sided alternative is desired, $t_K(\alpha)$ should be used. To have a power of $1 - \beta$, the required sample size is obtained by solving

$$1 - \beta = P\left\{ Z > t_K(\alpha/2) - \frac{\Delta}{\sigma\sqrt{2/n}} \right\}$$

Table 11.6.3 Upper 1% Points $t_k(\alpha)$ of the Distribution T_k

df <i>v</i>	<i>k</i> = Number of Dose Levels									
	1	2	3	4	5	6	7	8	9	10
5	3.36	3.50	3.55	3.57	3.59	3.60	3.60	3.61	3.61	3.61
6	3.14	3.26	3.29	3.31	3.32	3.33	3.34	3.34	3.34	3.35
7	3.00	3.10	3.13	3.15	3.16	3.16	3.17	3.17	3.17	3.17
8	2.90	2.99	3.01	3.03	3.04	3.04	3.05	3.05	3.05	3.05
9	2.82	2.90	2.93	2.94	2.95	2.95	2.96	2.96	2.96	2.96
10	2.76	2.84	2.86	2.88	2.88	2.89	2.89	2.89	2.90	2.90
11	2.72	2.79	2.81	2.82	2.83	2.83	2.84	2.84	2.84	2.84
12	2.68	2.75	2.77	2.78	2.79	2.79	2.79	2.80	2.80	2.80
13	2.65	2.72	2.74	2.75	2.75	2.76	2.76	2.76	2.76	2.76
14	2.62	2.69	2.71	2.72	2.72	2.73	2.73	2.73	2.73	2.73
15	2.60	2.66	2.68	2.69	2.70	2.70	2.70	2.71	2.71	2.71
16	2.58	2.64	2.66	2.67	2.68	2.68	2.68	2.68	2.68	2.69
17	2.57	2.63	2.64	2.65	2.66	2.66	2.66	2.66	2.67	2.67
18	2.55	2.61	2.63	2.64	2.64	2.64	2.65	2.65	2.65	2.65
19	2.54	2.60	2.61	2.62	2.63	2.63	2.63	2.63	2.63	2.63
20	2.53	2.58	2.60	2.61	2.61	2.62	2.62	2.62	2.62	2.62
22	2.51	2.56	2.58	2.59	2.59	2.59	2.60	2.60	2.60	2.60
24	2.49	2.55	2.56	2.57	2.57	2.57	2.58	2.58	2.58	2.58
26	2.48	2.53	2.55	2.55	2.56	2.56	2.56	2.56	2.56	2.56
28	2.47	2.52	2.53	2.54	2.54	2.55	2.55	2.55	2.55	2.55
30	2.46	2.51	2.52	2.53	2.53	2.54	2.54	2.54	2.54	2.54
35	2.44	2.49	2.50	2.51	2.51	2.51	2.51	2.52	2.52	2.52
40	2.42	2.47	2.48	2.49	2.49	2.50	2.50	2.50	2.50	2.50
60	2.39	2.43	2.45	2.45	2.46	2.46	2.46	2.46	2.46	2.46
120	2.36	2.40	2.41	2.42	2.42	2.42	2.42	2.42	2.42	2.43
∞	2.326	2.366	2.377	2.382	2.385	2.386	2.387	2.388	2.389	2.389

Source: Williams (1971).

and hence

$$-Z(\beta) = t_K(\alpha/2) - \frac{\Delta}{\sigma\sqrt{2/n}}.$$

Thus we have

$$n = \frac{2\sigma^2[t_K(\alpha/2) + Z(\beta)]^2}{\Delta^2}. \quad (11.6.3)$$

Example 11.6.1 To illustrate sample size calculation based on Williams's test, consider a dose-response study on three doses of an active drug in determining a minimum effective dose. Suppose that the standard deviation of the response variable is 45 and that the clinically meaningful difference is 25. By Table 11.6.2, we have $t_3(0.025)=2.032$. Then, by (11.6.3), we have

$$n = \frac{2(45)^2(2.032 + 0.842)^2}{(25)^2}$$

$$= 54.$$

Table 11.6.4 Upper 0.5% Points $t_k(\alpha)$ of the Distribution T_k

df v	k = Number of Dose Levels						
	2	3	4	5	6	8	10
5	4.179	4.229	4.255	4.270	4.279	4.292	4.299
6	3.825	3.864	3.883	3.895	3.902	3.912	3.197
7	3.599	3.631	3.647	3.657	3.663	3.670	3.674
8	3.443	3.471	3.484	3.492	3.497	3.504	3.507
9	3.329	3.354	3.366	3.373	3.377	3.383	3.886
10	3.242	3.265	3.275	3.281	3.286	3.290	3.293
11	3.173	3.194	3.204	3.210	3.214	3.218	3.221
12	3.118	3.138	3.147	3.152	3.156	3.160	3.162
13	3.073	3.091	3.100	3.105	3.108	3.112	3.114
14	3.035	3.052	3.060	3.065	3.068	3.072	3.074
15	3.003	3.019	3.027	3.031	3.034	3.037	3.039
16	2.957	2.991	2.998	3.002	3.005	3.008	3.010
17	2.951	2.966	2.973	2.977	2.980	2.938	2.984
18	2.929	2.944	2.951	2.955	2.958	2.960	2.962
19	2.911	2.925	2.932	2.936	2.938	2.941	2.942
20	2.894	2.903	2.915	2.918	2.920	2.923	2.925
22	2.866	2.879	2.885	2.889	2.891	2.893	2.895
24	2.842	2.855	2.861	2.864	2.866	2.869	2.870
26	2.823	2.835	2.841	2.844	2.846	2.848	2.850
28	2.806	2.819	2.824	2.827	2.829	2.831	2.832
30	2.792	2.804	2.809	2.812	2.814	2.816	2.817
35	2.764	2.776	2.781	2.783	2.785	2.787	2.788
40	2.744	2.755	2.759	2.762	2.764	2.765	2.766
60	2.697	2.707	2.711	2.713	2.715	2.716	2.717
120	2.651	2.660	2.664	2.666	2.667	2.669	2.669
∞	2.607	2.615	2.618	2.620	2.621	2.623	2.623

Source: Williams (1972).

Thus 54 subjects per treatment group are needed in order to have an 80% power for determining the minimum effective dose for the subjects under study.

11.7 CROSSOVER DESIGNS

Point Hypotheses for Equality

Let Y_{ijk} be the response of the i th subject in the k th sequence at the j th period. Then the following model without consideration of unequal carryover effects can be used to describe a standard two-sequence, two-period crossover design:

$$Y_{ijk} = \mu + S_{ik} + P_j + T_{(j,k)} + e_{ijk}, \quad (11.7.1)$$

where i (subject) = 1, 2, ..., n_k , j (period), k (sequence) = 1, 2. In model (11.7.1), μ is the overall mean, S_{ik} is the random effect of the i th subject in the k th sequence, P_j is the fixed

effect of the j th period, $T_{(j,k)}$ is the direct fixed effect of the treatment administered at period j in sequence k , namely

$$T_{(j,k)} = \begin{cases} \text{Placebo} & \text{if } k=j, \\ \text{Test Drug} & \text{if } k \neq j, k=1, 2, j=1, 2, \end{cases}$$

and e_{ijk} is the within-subject random error in observing Y_{ijk} . For model (11.7.1) it is assumed that $\{S_{ik}\}$ are independently and identically distributed with mean 0 and variance σ_S^2 and that $\{e_{ijk}\}$ are independently distributed with mean 0 and variance σ_e^2 . $\{S_{ik}\}$ and $\{e_{ijk}\}$ are assumed to be mutually independent. Let us test the following hypotheses:

$$\begin{aligned} H_0: \mu_T &= \mu_P, \\ \text{vs. } H_a: \mu_T &\neq \mu_P. \end{aligned} \quad (11.7.2)$$

Under model (11.7.1), we can consider period differences for each subject within each sequence which are defined as

$$d_{ik} = \frac{1}{2} (Y_{i2k} - Y_{i1k}),$$

where $i = 1, \dots, n_k$; $k = 1, 2$. Then a test for hypotheses (11.7.2) can be obtained based on a two-sample t statistic as follows:

$$T_d = \frac{\bar{Y}_T - \bar{Y}_P}{\hat{\sigma}_d \sqrt{(1/n_1) + (1/n_2)}},$$

where

$$\begin{aligned} \bar{Y}_T &= \frac{1}{2} (\bar{Y}_{.21} + \bar{Y}_{.12}), \\ \bar{Y}_P &= \frac{1}{2} (\bar{Y}_{.11} + \bar{Y}_{.22}), \\ \hat{\sigma}_d^2 &= \frac{1}{n_1 + n_2 - 2} \sum_{k=1}^2 \sum_{i=1}^{n_k} (d_{ik} - \bar{d}_{.k})^2, \end{aligned}$$

and

$$\bar{Y}_{jk} = \frac{1}{n_k} \sum_{i=1}^{n_k} Y_{ijk},$$

$$\bar{d}_{.k} = \frac{1}{n_k} \sum_{i=1}^{n_k} d_{ik}.$$

Under the null hypothesis (11.7.2), T_d follows a t distribution with $n_1 + n_2 - 2$ degrees of freedom. We can reject the null hypothesis of (11.7.2) if

$$|T_d| > t(\alpha/2, n_1 + n_2 - 2).$$

Under the alternative hypothesis that $\mu_T = \mu_P + \Delta$, the power of the test T_d can be similarly evaluated. In the interest of balance, we assume that $n_1 = n_2 = n$; that is, each sequence will be allocated the same number of subjects at random. As a result, the sample

size per sequence for testing the hypotheses of equality (11.7.2) can be determined by the formula

$$n \geq \frac{2\sigma_d^2[t(\alpha/2, 2n - 2) + t(\beta, 2n - 2)]^2}{\Delta^2}, \quad (11.7.3)$$

where σ_d^2 can be estimated from previous studies and Δ is the clinically meaningful difference which we want to detect. If we need to have a power of 80% for detection of a difference of at least 20% of the unknown placebo mean, then (11.7.3) can be simplified as

$$n \geq [t(\alpha/2, 2n - 2) + t(\beta, 2n - 2)]^2 \left[\frac{CV}{20} \right]^2, \quad (11.7.4)$$

where

$$CV = \frac{\sqrt{2}\sigma_d}{\mu_p} \times 100\%.$$

Since $(2n - 2)$ in (11.7.3) and (11.7.4) are unknown, a numerical iterative procedure is required to solve for n .

Example 11.7.1 To illustrate (11.7.3), consider the problem of determining the sample size for a clinical trial that compares a test drug with a placebo under a standard two-sequence, two-period crossover design (Chow and Liu, 2000). For this trial, the sponsor is interested in having an 80% power for detection of a 20% difference between the test drug and the placebo. Based on the results from previous studies, it is estimated that the reference mean is 82.559 with a CV of 15.66. Thus (11.7.4) can be applied to determine the sample size per sequence. For iterative purposes we first guess that $n = 9$. This gives degrees of freedom $2(n - 2) = 16$, $t(0.025, 16) = 2.12$, and $t(0.2, 16) = 0.865$. By (11.7.4),

$$\begin{aligned} n &= (2.12 + 0.865)^2 \left[\frac{15.66}{20} \right]^2 \\ &= 5.5 \approx 6. \end{aligned}$$

We then start with $n = 6$ and repeat the same calculation, which gives $n = 5.9 \approx 6$ which is very close to the previous solution. Therefore we conclude that a total of $N = 2n = 12$ subjects are required to provide an 80% power for detection of a 20% difference of the reference mean at the 5% level of significance.

Interval Hypotheses for Equivalence

As pointed out by Chow and Liu (2000), the power approach for sample size determination based on the hypothesis of equality (11.7.2) is not statistically valid in assessing *equivalence* between treatments. For the assessment of equivalence between treatments under the standard two-sequence, two-period crossover design, it is suggested that the following interval hypotheses be tested:

$$\begin{aligned} H_0: \mu_T - \mu_P &\leq \theta_L \quad \text{or} \quad \mu_T - \mu_P \geq \theta_U, \\ \text{vs. } H_a: \theta_L < \mu_T - \mu_P < \theta_U, \end{aligned} \quad (11.7.5)$$

where θ_L and θ_U are some clinically meaningful limits for equivalence. The concept of interval hypotheses is to show equivalence by rejecting the null hypothesis of inequivalence. The above hypotheses can be decomposed into two sets of one-sided hypotheses:

$$\begin{aligned} H_{01}: \mu_T - \mu_P &\leq \theta_L, \\ \text{vs. } H_{a1}: \mu_T - \mu_P &> \theta_L, \end{aligned}$$

and

$$\begin{aligned} H_{02}: \mu_T - \mu_P &\geq \theta_U, \\ \text{vs. } H_{a2}: \mu_T - \mu_P &< \theta_U. \end{aligned}$$

Under model (11.7.1), Schuirmann (1987) proposes two one-sided test procedures for the above two one-sided hypotheses. We can reject the null hypothesis of inequivalence if

$$T_L = \frac{\bar{Y}_T - \bar{Y}_P - \theta_L}{\hat{\sigma}_d \sqrt{(1/n_1) + (1/n_2)}} > t(\alpha, n_1 + n_2 - 2)$$

and

$$T_U = \frac{\bar{Y}_T - \bar{Y}_P - \theta_U}{\hat{\sigma}_d \sqrt{(1/n_1) + (1/n_2)}} < -t(\alpha, n_1 + n_2 - 2).$$

Let $\theta = \mu_T - \mu_P$ and $\phi_S(\theta)$ be the power of Schuirmann's two one-sided tests at θ . Assuming that $n_1 = n_2 = n$, the power at $\theta = 0$ is given by

$$\begin{aligned} 1 - \beta &= \phi_S(0) \\ &= P\left\{ \frac{-\Delta}{\hat{\sigma}_d \sqrt{2/n}} + t(\alpha, 2n - 2) < \frac{Y}{\hat{\sigma}_d \sqrt{2/n}} < \frac{\Delta}{\hat{\sigma}_d \sqrt{2/n}} - t(\alpha, 2n - 2) \right\}, \quad (11.7.6) \end{aligned}$$

where $Y = \bar{Y}_T - \bar{Y}_P$. Since a central t distribution is symmetric about 0, the lower and upper endpoints of (11.7.6) are also symmetric about 0:

$$\frac{-\Delta}{\hat{\sigma}_d \sqrt{2/n}} + t(\alpha, 2n - 2) = - \left\{ \frac{\Delta}{\hat{\sigma}_d \sqrt{2/n}} - t(\alpha, 2n - 2) \right\}.$$

Therefore, $\phi_S(0) \geq 1 - \beta$ implies that

$$\left| \frac{\Delta}{\hat{\sigma}_d \sqrt{2/n}} - t(\alpha, 2n - 2) \right| \geq t(\beta/2, 2n - 2),$$

or that

$$n(\theta = 0) \geq 2[t(\alpha, 2n - 2) + t(\beta/2, 2n - 2)]^2 \left[\frac{\hat{\sigma}_d}{\Delta} \right]^2 \quad (11.7.7)$$

If we must have an 80% power for detection of a 20% difference of placebo mean, then (11.7.7) becomes

$$n(\theta = 0) \geq [t(\alpha, 2n - 2) + t(\beta/2, 2n - 2)]^2 \left[\frac{CV}{20} \right]^2, \quad (11.7.8)$$

We will now consider the case where $\theta \neq 0$. Since the power curves of Schuirmann's two one-sided test procedures are symmetric about zero (Phillips, 1990), we will only consider the case where $0 < \theta = \theta_0 < \Delta$. In this case, the statistic

$$\frac{Y - \theta_0}{\hat{\sigma}_d \sqrt{2/n}}$$

has a central t distribution with $2n - 2$ degrees of freedom. The power of Schuirmann's two one-sided test procedures can be evaluated at θ_0 , which is given by

$$1 - \beta = \phi_S(\theta_0) \\ = P \left\{ \frac{-\Delta - \theta_0}{\hat{\sigma}_d \sqrt{2/n}} + t(\alpha, 2n - 2) < \frac{Y - \theta_0}{\hat{\sigma}_d \sqrt{2/n}} < \frac{\Delta - \theta_0}{\hat{\sigma}_d \sqrt{2/n}} - t(\alpha, 2n - 2) \right\}. \quad (11.7.9)$$

Note that unlike the case where $\theta = 0$, the lower and upper endpoints of (11.7.9) are not symmetric about 0. Therefore, as indicated by Chow and Liu (2000), if we choose

$$\frac{\Delta - \theta_0}{\hat{\sigma}_d \sqrt{2/n}} - t(\alpha, 2n - 2) = t(\beta/2, 2n - 2),$$

then the resultant sample size may be too large to be of practical interest, and the power may be more than we need. As an alternative, Chow and Liu (2000) consider the inequality for obtaining an approximate formula for n

$$\phi_S(\theta_0) \leq P \left\{ \frac{Y - \theta_0}{\hat{\sigma}_d \sqrt{2/n}} < \frac{\Delta - \theta_0}{\hat{\sigma}_d \sqrt{2/n}} - t(\alpha, 2n - 2) \right\}.$$

As a result, $\phi_S(\theta_0) \geq 1 - \beta$ gives

$$\frac{\Delta - \theta_0}{\hat{\sigma}_d \sqrt{2/n}} - t(\alpha, 2n - 2) = t(\beta, 2n - 2),$$

or

$$n(\theta_0) \geq 2[t(\alpha, 2n - 2) + t(\beta, 2n - 2)]^2 \left[\frac{\hat{\sigma}_d}{\Delta - \theta_0} \right]^2. \quad (11.7.10)$$

Similarly, if we must have an 80% power for detection of a 20% difference of placebo mean, then (11.7.10) becomes

$$n(\theta_0) \geq [t(\alpha, 2n - 2) + t(\beta, 2n - 2)]^2 \left[\frac{CV}{20 - \theta'_0} \right]^2, \quad (11.7.11)$$

where

$$\theta'_0 = 100 \times \frac{\theta_0}{\mu_P}.$$

Example 11.7.2 We continue the previous example to illustrate the computation of the sample size in achieving an 80% power when $\theta'_0 = 5\%$ and $\Delta = 20\%$. From the previous example, the estimated CV is given by 15.66%. If our initial guess of n is 6, then

$$\begin{aligned} t(\alpha, 2n - 2) &= t(0.05, 10) = 1.812, \\ t(\beta, 2n - 2) &= t(0.20, 10) = 1.879. \end{aligned}$$

Thus

$$\begin{aligned} n &= (1.812 + 0.879)^2 \left[\frac{15.66}{20 - 5} \right]^2 \\ &= 7.9 \approx 8. \end{aligned}$$

We then use $n = 8$ as the initial value for the next enumeration. Since $t(0.05, 14) = 1.761$ and $t(0.20, 14) = 0.868$, (11.7.11) gives

$$\begin{aligned} n &= (1.761 + 0.868)^2 \left[\frac{15.66}{20 - 5} \right]^2 \\ &= 7.9 \approx 8. \end{aligned}$$

Since the last two enumerations generate the same required sample size per sequence a total of 16 subjects would be required to achieve an 80% power.

Table 11.7.1 presents the required total sample sizes necessary to achieve either an 80% or a 90% power for θ from 0 to 15% by increments of 5% as well as CVs from 10% to 40% by increments of 2% (Liu and Chow, 1992).

Higher-Order Crossover Designs

As was demonstrated in Chapter 5, the standard two-sequence, two-period crossover design is not useful in the presence of carryover effects. In addition it does not provide independent estimates of intrasubject variabilities. To account for these disadvantages, in practice, it is of interest to consider a higher-order crossover design (Chow and Liu, 1992, 2000; Chow and Wang, 2001). A higher-order crossover design is defined as a crossover design in which either the number of periods or the number of sequences is greater than the number of treatments to be compared. The most commonly used higher-order crossover designs are Balaam's design, the two-sequence dual design, and four-period designs with two or four sequences. For a higher-order crossover design, the general model given in (11.7.1) can be used with the appropriate indexes on j and k . Table 11.7.2 summarizes these higher-order crossover designs.

For a given higher-order crossover design, the sample size is similarly determined based on either the point hypotheses for equality or the interval hypotheses for equivalence under model (11.7.1). Let us consider the sample size determined by the interval hypotheses (11.7.5). Let $n_i = n$ be the number of subjects in sequence i of a higher-order crossover design, and let F_v denote the cumulative distribution function of the t distribution with v degrees of freedom. Then it can be verified that the power of Schuirmann's two one-sided tests at the α level of significance for the m th design is given by

$$\begin{aligned} \phi_m(\theta) &= F_{v_m} \left(\frac{\Delta - \theta}{CV \sqrt{b_m/n}} - t(\alpha, v_m) \right) \\ &\quad - F_{v_m} \left(t(\alpha, v_m) - \frac{\Delta + \theta}{CV \sqrt{b_m/n}} \right) \end{aligned}$$

Table 11.7.1 Sample Sizes for Schuirmann's Two One-Sided Test Procedures at $\Delta = 0.2$ and the 5% Nominal Level in 2×2 Standard Crossover Design

Power	CV(%)	$\theta = \mu_T - \mu_R$			
		0%	5%	10%	15%
80%	10	8	8	16	52
	12	8	10	20	74
	14	10	14	26	100
	16	14	16	34	126
	18	16	20	42	162
	20	20	24	52	200
	22	24	28	62	242
	24	28	34	74	288
	26	32	40	86	336
	28	36	46	100	390
	30	40	52	114	448
	32	46	58	128	508
	34	52	66	146	574
	36	58	74	162	644
	38	64	82	180	716
	40	70	90	200	794
90%	10	10	10	20	70
	12	10	14	28	100
	14	14	18	36	136
	16	16	22	46	178
	18	20	28	58	224
	20	24	32	70	276
	22	28	40	86	334
	24	34	46	100	396
	26	40	54	118	466
	28	44	62	136	540
	30	52	70	156	618
	32	58	80	178	704
	34	66	90	200	794
	36	72	100	224	890
	38	80	112	250	992
	40	90	124	276	1098

Source: Liu and Chow (1992).

for $m=1$ (Balaam design), 2 (two-sequence dual design), 3 (four-period design with two sequences), and 4 (four-period design with four sequences), where

$$v_1 = 4n - 3, \quad v_2 = 4n - 4, \quad v_3 = 6n - 5, \quad v_4 = 12n - 5;$$

$$b_1 = 2, \quad b_2 = \frac{3}{4}, \quad b_3 = \frac{11}{20}, \quad b_4 = \frac{1}{4}.$$

Table 11.7.2 Commonly Used Higher-Order Crossover Designs

Design	Sequence	Period	Design
1	4	2	Balam design
2	2	3	Two-sequence dual design
3	2	4	Four-period design with two sequences
4	4	4	Four-period design with four sequences

Hence the formula of n required to achieve a $1 - \beta$ power at the α level of significance for the m th design when $\theta = 0$ is given by

$$n \geq b_m [t(\alpha, v_m) + t(\beta/2, v_m)]^2 \left[\frac{CV}{\Delta} \right]^2, \quad (11.7.12)$$

and if $\theta = \theta_0 > 0$, the approximate formula for n is given by

$$n(\theta_0) \geq b_m [t(\alpha, v_m) + t(\beta, v_m)]^2 \left[\frac{CV}{\Delta - \theta} \right]^2 \quad (11.7.13)$$

for $m = 1, 2, 3$, and 4.

Note that Tables 11.7.3 through 11.7.6 give the required total number of subjects N_m for each dosing (m) to achieve either an 80% or a 90% power for θ from 0 to 15% by increments of 5% as well as CVs from 10% to 40% by increments of 2%, where

$$N_1 = 4n, \quad N_2 = 2n, \quad N_3 = 2n, \quad \text{and} \quad N_4 = 4n.$$

11.8 EQUIVALENCE AND NONINFERIORITY TRIALS

As indicated in Chapter 7, equivalence or noninferiority trials have been widely employed for evaluation of new treatments. These new treatments may be developed for less invasive or easier administration (better safety profiles) or for lower cost. Therefore, it is very important to verify that these new treatments can still maintain *similar* efficacy as compared to the standard treatment. In here, similarity can be defined as two-sided equivalence or one-sided noninferiority. The controversies in design and analysis of an active equivalence/noninferiority trial and in selection of equivalence limits have been discussed in Chapter 7. Sample size determination is another controversial yet important issue in design of equivalence and noninferiority trials because it depends on the equivalence limits. The more stringent the equivalence limits, the larger the sample size required. In practice, the equivalence limits are usually chosen based on results observed from similar studies conducted previously. As a result, equivalence limits are in fact estimates, whose variability should be accounted for. In this section, however, statistical methods or procedures for sample size calculation based on binary or survival endpoints are reviewed under the assumption that the equivalence limits are predefined known constants. Note that the sample size required for a two-sided equivalence trial is generally larger than that required for a one-sided noninferiority trial, which is in turn larger than that for a superiority trial. More details can be found in the editorial by Ware and Antman (1997) for the COBALT study (1997) and Chow et al. (2003).

Table 11.7.3 Sample Sizes for Schuirmann's Two One-Sided Test Procedures at $\Delta = 0.2$ and the 5% Nominal Level in Balaam's Design

Power	CV(%)	θ			
		0%	5%	10%	15%
80%	10	20	24	52	200
	12	28	36	76	288
	14	36	48	100	392
	16	48	60	132	508
	18	60	76	164	644
	20	72	92	200	796
	22	88	108	244	960
	24	104	132	288	1144
	26	120	152	336	1340
	28	136	176	392	1556
	30	156	200	448	1784
	32	180	228	508	2028
	34	200	256	576	2292
	36	224	288	644	2568
	38	252	320	716	2860
	40	276	356	796	3168
90%	10	24	36	72	276
	12	36	48	104	400
	14	48	64	136	540
	16	60	80	180	704
	18	76	104	224	892
	20	92	124	276	1100
	22	108	152	336	1328
	24	128	180	400	1584
	26	152	208	468	1856
	28	172	244	540	2152
	30	200	276	620	2472
	32	224	316	704	2808
	36	284	400	892	3556
	38	316	444	992	3960
	40	352	492	1100	4388

11.8.1 Independent Binary Endpoints

There is a vast literature on statistical methodology for testing equivalence or noninferiority for binary outcomes (e.g., success/failure or yes/no) obtained from parallel group studies. For example, for asymptotic methods, see Dunnet and Gent (1977), Blackwelder (1982), Farrington and Manning (1990), Nam (1995), Dunnet and Gent (1996), Chen et al. (2000), and Chow et al. (2003). For exact methods, see Chan (1998), Rohmel and Mansmann (1999), Chan and Zhang (1999), Kang and Chen (2000), Chan (2003), and Chow et al. (2003b).

In this section, we first review the method for sample size determination based on the asymptotic method proposed by Farrington and Manning (1990). Suppose a clinical trial is conducted to compare a test treatment in n_T subjects to a standard treatment in n_C subjects. Let Y_T and Y_C be the number of subjects with a (success) response to their treatment,

Table 11.7.4 Sample Sizes for Schuirmann's Two One-Sided Test Procedures at $\Delta = 0.2$ and the 5% Nominal Level in Two-Sequence Dual Design

Power	CV(%)	θ			
		0%	5%	10%	15%
80%	10	6	6	12	38
	12	6	8	16	56
	14	8	10	20	74
	16	10	12	26	96
	18	12	16	32	122
	20	14	18	38	150
	22	18	22	46	182
	24	20	26	56	216
	26	24	30	64	252
	28	28	34	74	292
	30	30	38	86	336
	32	34	44	96	382
	34	38	50	108	430
	36	44	56	122	482
	38	48	62	136	538
	40	54	68	150	596
90%	10	6	8	14	54
	12	8	10	20	76
	14	10	14	28	102
	16	12	16	34	134
	18	16	20	44	168
	20	18	24	54	208
	22	22	30	64	250
	24	26	34	76	298
	26	30	40	88	350
	28	34	46	102	404
	30	38	54	118	464
	32	44	60	134	528
	34	48	68	150	596
	36	54	76	168	668
	38	60	84	188	744
	40	66	94	208	824

respectively. Denote P_T and P_C as the true unknown response rate of the test and standard treatments, respectively. The noninferiority hypothesis in (7.4.2) can be expressed for the proportion difference $\theta = P_T - P_C$ as

$$H_0: P_T - P_C \leq -\Delta \quad \text{vs.} \quad H_a: P_T - P_C > -\Delta, \quad (11.8.1)$$

where Δ is a prespecified equivalence limit of clinical importance.

Noninferiority of the test treatment, as compared to the standard treatment, is concluded at the α level of significance if

$$T_L = (p_T - p_C + \Delta)/\sqrt{v_0} > z(\alpha), \quad (11.8.2)$$

Table 11.7.5 Sample Sizes for Schuirmann's Two One-Sided Test Procedures at $\Delta = 0.2$ and the 5% Nominal Level in Four-Period Design With Two Sequences

Power	CV(%)	θ			
		0%	5%	10%	15%
80%	10	4	4	8	28
	12	6	6	12	40
	14	6	8	14	54
	16	8	10	18	72
	18	10	12	24	90
	20	12	14	28	110
	22	14	16	34	134
	24	16	18	40	158
	26	18	22	48	186
	28	20	26	54	214
	30	22	28	62	246
	32	26	32	72	280
	34	28	36	80	316
	36	32	40	90	354
	38	36	46	100	394
	40	40	50	110	436
90%	10	4	6	12	40
	12	6	8	16	56
	14	8	10	20	76
	16	10	12	26	98
	18	12	16	32	124
	20	14	18	40	152
	22	16	22	48	184
	24	18	26	56	218
	26	22	30	66	256
	28	24	34	76	296
	30	28	40	86	340
	32	32	44	98	388
	34	36	50	110	438
	36	40	56	124	490
	38	44	62	138	546
	40	50	68	152	604

where p_T and p_C are the observed sample response rates for the test and the standard treatments, respectively,

$$\begin{aligned} v_0 &= p'_T(1 - p'_T)/n_T + p'_C(1 - p'_C)/n_C \\ &= [p'_T(1 - p'_T) + p'_C(1 - p'_C)/r]/n_T, \end{aligned} \quad (11.8.3)$$

and p'_T and p'_C are the restricted maximum likelihood estimators (REML) of P_T and P_C obtained under the boundary point of the null hypothesis $P_T - P_C = -\Delta$, and $r = n_C/n_T$.

The RMLE p'_T is the unique solution in $(-\Delta, 1)$ of the following maximum likelihood equation:

$$ax^3 + bx^2 + cx + d = 0,$$

Table 11.7.6 Sample Sizes for Schuirmann's Two One-Sided Test Procedures at $\Delta = 0.2$ and the 5% Nominal Level in Four-Period Design With Four Sequences

Power	CV(%)	θ			
		0%	5%	10%	15%
80%	10	4	4	8	28
	12	4	8	12	40
	14	8	8	16	52
	16	8	8	20	64
	18	8	12	24	84
	20	12	12	28	100
	22	12	16	32	124
	24	16	20	40	144
	26	16	20	44	168
	28	20	24	52	196
	30	20	28	60	224
	32	24	32	64	256
	34	28	36	72	288
	36	32	40	84	324
	38	32	44	92	360
	40	36	48	100	400
90%	10	4	8	12	36
	12	8	8	16	52
	14	8	12	20	68
	16	8	12	24	92
	18	12	16	32	112
	20	12	16	36	140
	22	16	20	44	168
	24	20	24	52	200
	26	20	28	60	236
	28	24	32	68	272
	30	28	36	80	312
	32	32	40	92	352
	34	32	48	100	400
	36	36	52	112	448
	38	40	56	128	496
	40	44	64	140	552

where $a = 1 + r$, $b = -[1 + r + p_T + rp_C - \Delta(r + 2)]$, $c = \Delta^2 - \Delta(2p_T + r + 1) + p_T + rp_C$, and $d = p_T\Delta(1 - \Delta)$. The solution is given by

$$p'_T = 2u \cos(w) - b/3a \quad \text{and} \quad p'_C = p'_T + \Delta,$$

where

$$w = [\pi + \cos^{-1}(v/u^3)]/3, v = b^3/(3a)^3 - bc/(6a^2) + d/2a \quad \text{and}$$

$$u = \operatorname{sgn}(v)[b^2/(3a)^2 - c/3a]^{1/2}.$$

The asymptotic formula for sample size for achieving a desired power of $1 - \beta$ for testing the hypotheses in (11.8.1) at the α level of significance level is then given by

$$\begin{aligned} n_T \geq & \{Z(\alpha) \sqrt{[p'_T(1 - p'_T) + p'_C(1 - p'_C)/r] + Z(\beta) \sqrt{[P_T(1 - P_T)}} \\ & + P_C(1 - P_C)/r\}]^2/[P_C - P_T - \Delta]^2. \end{aligned} \quad (11.8.4)$$

In many trials, however, the primary endpoint for evaluation of efficacy for binary outcomes is not the difference in proportion but the relative risk (i.e., the ratio of the proportion of the subject with the predefined risk of the test treatment to the standard treatment). The risk could be occurrence of a certain disease or death. The noninferiority hypothesis based on relative risk can be formulated as

$$H_0: R \geq R_0 \quad \text{vs.} \quad H_a: R < R_0, \quad (11.8.5)$$

where $R = P_T/P_C$, the relative risk of the test treatment to the standard treatment, and R_0 is some prespecified tolerance limit. Based on the relative risk, the noninferiority of the test treatment, as compared to the standard treatment, is concluded at the α level of significance if

$$T_L = (p_T - R_0 p_C)/\sqrt{w_0} < Z(\alpha), \quad (11.8.6)$$

where p_T and p_C are the sample observed response rates for the test and the standard treatments, respectively,

$$\begin{aligned} w_0 &= p_T^*(1 - p_T^*)/N_T + p_C^*(1 - p_C^*)/n_C \\ &= [p_T^*(1 - p_T^*) + (R_0/r)p_C^*(1 - p_C^*)/r]/n_T, \end{aligned} \quad (11.8.7)$$

and p_T^* and p_C^* are the REML of P_T and P_C obtained under the boundary point of the null hypothesis $P_T = R_0 P_C$, and $r = n_C/n_T$. The RMLE p_T^* is the unique solution in $(0, 1)$ of the following maximum likelihood equation:

$$ax^2 + bx + c = 0,$$

where $a = 1 + r$, $b = -[R_0(1 + rp_C) + r + p_T]$, and $c = R_0(p_T + rp_C)$. The solution is given by

$$p_T^* = [-b - (b^2 - 4ac)^{1/2}]/2a, \quad \text{and} \quad p_C^* = p_T^*/R_0. \quad (11.8.8)$$

For the relative risk, the asymptotic formula for sample size for achieving a desired power of $1 - \beta$ for testing the hypotheses in (11.8.5) at the α level of significance is then given by

$$\begin{aligned} n_T \geq & \{Z(\alpha) \sqrt{[p_T^*(1 - p_T^*) + (R_0/r)p_C^*(1 - p_C^*)/r] + Z(\beta) \sqrt{[P_T(1 - P_T)}} \\ & + (R_0/r)P_C(1 - P_C)/r\}]^2/[P_T - R_0 P_C]^2. \end{aligned} \quad (11.8.9)$$

Table 11.8.1 provides the required sample sizes for achieving a 90% power for testing the noninferiority hypothesis at the 5% level of significance based on independent binary

Table 11.8.1 Sample Sizes¹ per Group with Equal Allocation for Noninferiority Hypothesis Based on Independent Binary Outcomes for Two Group Parallel Design

P_T	P_C	Δ	n_T	R_0	n_T
0.1	0.1	-0.2	46	0.1	46
0.2	0.1	-0.1	57	0.5	57
0.05	0.05	-0.2	32	0.1	96
0.1	0.05	-0.05	128	0.5	126
0.01	0.01	-0.2	18	0.1	501
0.02	0.01	-0.01	695	0.5	677

¹ Sample size for a 90% power at the 5% significance level.

Summarized from Farrington and Manning (1990).

outcomes for various combinations of P_T , P_C , Δ , and R_0 . The sample size formulas given in (11.8.4) and (11.8.9) are derived from the asymptotic procedures for testing the noninferiority hypothesis that are generally valid when the sample size is large and P_T and P_C are in the range between 0.2 and 0.8. These asymptotic procedures may inflate type I error rate when sample size is small and P_T and P_C approach to either 0 or 1. Under these circumstances, exact procedures based on the exact (or permutational) distributions of the test statistics, which provide a valid inference, should be used. However, the calculation of p -value, power, and sample size determination require extensive computation because of the iterative nature of the exact procedures. For details on the sample size determination based on the exact tests for evaluation of noninferiority between two treatments with independent binary endpoints, see Chan (2003).

11.8.2 Paired Binary Endpoints

Evaluation of noninferiority or equivalence between the new and standard treatments sometimes is based on paired binary outcomes. For example, in comparing diagnostic efficacy of a new test noninvasive diagnostic procedure to a standard invasive diagnostic procedure, the patients will first receive the test diagnostic procedure followed by the standard invasive diagnostic procedure. For this type of study, two binary outcomes are observed in each patient. Because they are observed from the same patient, these two binary endpoints are correlated. Recently, statistical procedures have been proposed for evaluation of noninferiority or equivalence based on paired binary endpoints (Lu and Bean, 1995; Nam, 1997; Tango, 1998; Lachenburg and Lynch, 1998; Lui and Cumberland, 2001; Liu et al., 2002). In what follows, we introduce the sample size formula based on the asymptotic method proposed by Liu et al. (2002).

Let Y_{0j} , Y_{1j} , $j = 0$ (no or failure), and 1 (yes or success) be the binary responses representing the diagnostic results for the test and the standard procedures, respectively. Let y_{00} , y_{01} , y_{10} , and y_{11} be the observed numbers of pairs (0,0), (0,1), (1,0), and (1,1). Table 11.8.2 provides the 2×2 table for the four outcomes and probabilities. Note that $P_T = P_{11} + P_{10}$ and $P_C = P_{11} + P_{01}$. The noninferiority hypothesis of (11.8.1) can be expressed in terms of disconcordant probabilities as follows:

$$H_0: P_{10} - P_{01} \leq -\Delta \quad \text{vs.} \quad H_a: P_{10} - P_{01} > -\Delta. \quad (11.8.10)$$

Table 11.8.2 The Four Outcomes and Probabilities

		Diagnosis for Standard Procedure		
Diagnosis for Test Procedure		1(Yes)	2(No)	Total
1(Yes)		$y_{11}(P_{11})$	$y_{10}(P_{10})$	$y_T(P_T)$
0(No)		$y_{01}(P_{01})$	$y_{00}(P_{00})$	$n - y_T(1 - P_T)$
		$y_C(P_C)$	$n - y_C(1 - P_C)$	n

Following Nam (1997) and Liu et al. (2002), the noninferiority of the test procedure to the standard procedure is concluded at the α level of significance if

$$T_L = (p_{10} - p_{01} + \Delta)/\sqrt{s_0} > Z(\alpha), \quad (11.8.11)$$

where p_{10} and p_{01} are the observed sample response rates for the test and standard respectively,

$$s_0 = (p'_{10} + p'_{01}) - \Delta^2, \quad (11.8.12)$$

and p'_{10} and p'_{01} are the maximum likelihood estimators of P_{10} and P_{01} evaluated at the boundary point of the null hypothesis $P_{10} - P_{01} = -\Delta$, which are given by

$$p'_{01} = [-a + (a^2 - 8b)^{1/2}]/4, \quad p'_{10} = p'_{01} - \Delta,$$

where $a = -(p_{10} - p_{01})(1 - \Delta) - 2(p_{01} + \Delta)$ and $b = \Delta(1 + \Delta)p_{01}$.

The asymptotic formula for sample size for achieving a desired power of $1 - \beta$ for testing the hypotheses in (11.8.10) at the α level of significance when $P_T - P_C = 0$ is then given by

$$n \geq 2P_{10}\{(Z(\alpha)/g) + Z(\beta)\}/\Delta^2, \quad (11.8.13)$$

where $g = [2P_{10}/(2p'_{10} - \Delta - \Delta^2)]^{1/2}$.

Table 11.8.2 provides the required sample sizes for achieving an 80% power for testing the noninferiority hypothesis at the 5% level of significance based on paired binary outcomes for various values of Δ and $P_{01} = P_{10}$. Although the test procedure proposed by Liu et al. (2002) is a generalization of the McNemar's test, its interpretation is very different. For the McNemar's test for detection of existence of difference between P_T and P_C , the McNemar's test takes into account only the information of discordant pairs y_{10} and y_{01} . The interpretation of the results of 10 discordant pairs in a sample of 20 subjects is very different from the same number of discordant pairs in a sample of 10,000 subjects in testing the equivalence/noninferiority hypothesis. Therefore, concordant pairs provide information on the consistency between the two treatments and should not be ignored for the construction of the statistical procedures and confidence interval in testing the equivalence or noninferiority hypothesis. When sample size is small or the number of discordant pairs is small, the asymptotic tests are not reliable for controlling the type I error rate. Hsueh et al. (2001) proposed unconditional exact tests for testing the equivalence or noninferiority hypotheses based on paired binary endpoints with the corresponding confidence intervals. However, computation of the unconditional exact confidence interval is intensive. For recent developments on exact unconditional tests for paired binary endpoints, see Berger and Sidik (2003).

Table 11.8.3 Sample Sizes¹ per Group for Noninferiority Hypothesis Based on Paired Binary Outcomes when $P_{01} = P_{10}$

P_{01}	Δ	n
0.05	0.05	350
0.10	0.05	644
0.15	0.05	949
0.15	0.10	242
0.15	0.15	111
0.20	0.05	1259
0.20	0.10	317
0.20	0.15	142
0.20	0.20	81
0.30	0.20	116

¹ Sample size for a 80% power at the 2.5% significance level.

Summarized from Liu et al. (2002).

11.8.3 Independent Censored Endpoints

As mentioned in Chapter 10, the primary endpoint for some clinical trials are censored data, which are defined as the time from randomization to the occurrence of a predefined event, such as death, eradication of an infection caused by a certain microorganism, or progression of the disease. Recently, an increasing number of trials was conducted to evaluate noninferiority or equivalence of new cytotoxic chemotherapy for cancers to the standard treatment because of a reduced toxicity and a better quality of life offered by the new treatment. Therefore, it is very important to verify that the survival provided the new treatment is no worse than or equivalent to the standard treatment if it is not superior to the standard treatment. Com-Nougue et al. (1993) proposed a test for a one-sided hypothesis in the relative risk to establish the noninferiority of the new treatment. On the other hand, under the assumption of proportional hazards, Wellek (1993) derived an asymptotic uniformly most power (UMP) test for equivalence of survival functions between two treatments.

For two independent samples, let $S_C(\cdot)$ and $S_T(\cdot)$ be survival functions for the test and standard treatments, respectively. In addition, we assume that $S_C(\cdot)$ and $S_T(\cdot)$ are from the same proportional hazard model:

$$S_T(t) = [S_C(t)]^\lambda \quad \text{for all } t > 0 \text{ and some } \lambda,$$

where $\lambda = e^\theta$.

The equivalence hypothesis in (7.4.1) can be expressed as the difference between two survival functions as:

$$H_0: \sup |S_C(y) - S_T(y)| \geq \delta \quad \text{vs.} \quad H_a: \sup |S_C(y) - S_T(y)| < \delta, \quad (11.8.14)$$

for some $\delta > 0$.

Wellek (1993) showed that the hypotheses in (11.8.14) could be reformulated in terms of θ as

$$H_0: \sup |\theta| \geq \theta^* \quad \text{vs.} \quad H_a: \sup |\theta| < \theta^*, \quad (11.8.15)$$

where $\exp[\theta^*/(1 - e^{\theta^*})] - \exp[\theta^*e^{\theta^*}/(1 - e^{\theta^*})] = \delta$.

Let r_{jv} be the number of subjects at risk in the v th sample at the j th smallest failure time $y_{(j)}$, d_{+v} be total number of failures in the v th sample, and d_{j+} be the total number of failures at the j th smallest failure time. The maximum likelihood estimator of θ satisfies

$$\sum d_{j+} [r_{j2}e^\theta / (r_{j1} + r_{j2}e^\theta)] = d_{+2}, \quad (11.8.16)$$

and the observed information at the MLE is given as

$$I(\hat{\theta}) = \sum d_{j+} r_{j1} r_{j2} e^\theta / (r_{j1} + r_{j2} e^\theta)^2. \quad (11.8.17)$$

Wellek (1993) proposed an asymptotic uniformly most powerful (UMP) level α test with the rejection region

$$|\hat{\theta}| \sqrt{I(\hat{\theta})} < \chi[\alpha, \theta^* \sqrt{I(\hat{\theta})}], \quad (11.8.18)$$

where $\chi[\alpha, q]$ is the square root of the α th upper quantile of a noncentral chi-square distribution with 1 degree of freedom and noncentrality parameter q .

It is easy to see that the hypotheses of (11.8.15) can be decomposed into two one-sided hypotheses as

$$H_{0U}: \theta \leq -\theta^* \quad \text{vs.} \quad H_{aU}: \theta > -\theta^*$$

and

$$H_{0L}: \theta \geq \theta^* \quad \text{vs.} \quad H_{aL}: \theta < \theta^*.$$

The survival functions of the test and standard treatments are claimed equivalence with respect to equivalence limit δ at the α significant level if and only if

$$T_L = (\hat{\theta} - \theta^*) / \sqrt{I(\hat{\theta})} < -Z(\alpha)$$

and

$$T_U = (\hat{\theta} + \theta^*) / \sqrt{I(\hat{\theta})} > Z(\alpha). \quad (11.8.20)$$

Denote $1 - G_C(y)$ and $1 - G_T(y)$ as the distribution function of the censored time for the test and standard treatments, respectively, and $\rho = \lim_{n \rightarrow \infty} n_T/n$, where $n = n_T + n_C$.

The asymptotic formula for sample size for achieving a desired power of $1 - \beta$ for testing the hypotheses in (11.8.15) at the α level of significance at $\theta = 0$ based on the Wellek's test in (11.8.18) and based on the two one-sided tests' procedure in (11.8.20) are given by

$$\chi[\alpha, \sqrt{N}\theta^*/v(0)] > Z(\beta/2) \quad (11.8.21)$$

and

$$N = [v^2(0)/\theta^{*2}] [Z(\alpha) + Z(\beta/2)]^2, \quad (11.8.22)$$

respectively, where

$$\begin{aligned} 1/v^2(0) &= \int p(y)q(y)u(y)dy, \\ p(y) &= \rho G_T(y)/[\rho G_T(y) + (1 - \rho)G_C(y)] = 1 - q(y), \text{ and} \\ u(y) &= [\rho G_T(y) + (1 - \rho)G_C(y)]\{d[1 - S_C(y)]/dy\}. \end{aligned} \quad (11.8.23)$$

The noninferiority hypothesis for the survival functions is formulated as

$$H_0: \inf [S_T(y) - S_C(y)] \leq -\delta \quad \text{vs.} \quad H_a: \inf [S_T(y) - S_C(y)] > -\delta, \quad (11.8.24)$$

or equivalently,

$$H_{0L}: \theta \geq \theta^* \quad \text{vs.} \quad H_{aL}: \theta < \theta^*.$$

As a result, T_L in (11.8.20) can be employed in testing the noninferiority hypothesis in (11.8.24). The asymptotic sample size for achieving a desired power of $1 - \beta$ for testing the noninferiority hypothesis in (11.8.24) at the α level of significance at $\theta = 0$ is given by

$$N = [\nu^2(0)/\theta^{*2}][Z(\alpha) + Z(\beta)]^2 \quad (11.8.25)$$

Table 11.8.4 provides the empirical type I error rate and power through simulation for the Wellek's test and the two one-sided tests' procedure (TOST) and for the noninferiority test for various sample sizes at the 5% level of significance with equivalence limit $\delta = 0.10$ ($\theta^* = 0.2727$) when $\theta = 0$. It can be seen that from Table 11.8.4 that although the Wellek's is an asymptotic UMP test, its advantage over the two one-sided tests' procedure

Table 11.8.4 Empirical Size and Power for the Wellek's test and TOST for Equivalence and the Noninferiority Test at 5% Significance Level for $\delta = 0.10$ ($\theta^* = 0.2727$)

Distribution	Total Sample Size	Test	Size	Power
Lognormal	400	Wellek	0.0491	0.5800
		TOST	0.0483	0.5763
		NI	0.0490	0.7918
	500	Wellek	0.0497	0.7268
		TOST	0.0496	0.7266
		NI	0.0496	0.8629
	600	Wellek	0.0490	0.8254
		TOST	0.0490	0.8254
		NI	0.0490	0.9111
Exponential	700	Wellek	0.0502	0.8839
		TOST	0.0502	0.8839
		NI	0.0502	0.9425
	400	Wellek	0.0500	0.5609
		TOST	0.0496	0.5774
		NI	0.0498	0.7803
	500	Wellek	0.0512	0.7223
		TOST	0.0512	0.7214
		NI	0.0512	0.8564
	600	Wellek	0.0510	0.8121
		TOST	0.0510	0.8121
		NI	0.0510	0.9077
	700	Wellek	0.0475	0.8821
		TOST	0.0475	0.8821
		NI	0.0475	0.9426

NI = noninferiority one-sided test.

Table 11.8.5 Sample Size Required per arm with Equal Allocation for the Wellek's Test and TOST for Equivalence and the Noninferiority Test at the 5% Significance Level for $\delta = 0.15$, $S_c(5) = 0.55$, Censored Rate = 0.20

Power	Method	Log-normal Distribution			Exponential Distribution		
		$(\infty, *)^a$	(5.1) ^b	(5.2) ^c	$(\infty, *)^a$	(5.1) ^b	(5.2) ^c
70%	Wellek	107	302	233	107	271	224
	TOST	107	302	234	107	271	224
	NI	70	198	153	70	178	147
80%	Wellek	127	360	278	127	323	266
	TOST	127	360	278	127	323	266
	NI	92	260	201	92	234	192
90%	Wellek	161	454	351	161	408	336
	TOST	161	454	351	161	408	336
	NI	127	360	272	127	323	266

 $(\infty, *)^a$: infinity accrual and follow-up(5.1)^b: 5 years of uniform accrual and 1 additional year of follow-up(5.2)^c: 5 years of uniform accrual and 2 additional year of follow-up

NI=noninferiority one-sided test.

diminishes as the sample size increases. When sample is moderately large, the TOST procedure in (11.8.20) not only can control the type I error rate, but also it provides the same power as the Wellek's test without necessity for enumeration of noncentrality parameter. In fact, Hsueh et al. (2002) showed that when the total sample size is sufficiently large to provide a power at least 80%, the performance of the Wellek's test and the TOST procedure in terms of size and power are identical. As a result, the required sample sizes are identical for both methods as shown in Table 11.8.5.

11.9 MULTIPLE-STAGE DESIGN IN CANCER TRIALS

As was discussed in Section 6.6, in phase II cancer trials, it is undesirable to stop a study early when the treatment appears to be effective but desirable when the treatment seems to be ineffective. For this purpose, a multiple-stage design is often employed to determine whether an experimental treatment holds sufficient promise to warrant further testing (see, e.g., Fleming, 1982; Simon, 1989; Chang et al., 1987; Therneau et al., 1990). The concept of a multiple-stage design is to permit early stopping when a moderately long sequence of initial failures occurs. Chow et al. (2003b) provided tables for sample size calculation for two-stage designs such as minimax design, Simon's optimal two-stage design, and flexible two-stage design and three-stage designs such as optimal three-stage design and flexible design for single-arm phase II cancer trials.

Example 11.9.1 Suppose an investigator is interested in planning a clinical trial in patients with a specific carcinoma where standard therapy has a 20% response rate. Suppose further

that the test treatment is a combination of the standard therapy and a new agent. The investigator would like to determine whether the test treatment will achieve a response rate of 40%. Since the test treatment may not be effective, it is desirable to warrant early study termination. In this case, the optimal three-stage design described above is useful. We consider the optimal three-stage design for testing

$$H_0: p \leq 0.20 \quad \text{vs.} \quad H_1: p \geq 0.40$$

with $\alpha = \beta = 0.10$ which result in the following sample size allocation at each stage:

$$(0/8, 3/16, 11/42).$$

That is, at stage 1, 8 patients are to be tested. We would terminate the trial if no response is observed in the eight patients. If there are one or more responses, we continue to the next stage. At the second stage, eight more patients are treated. We stop the trial if fewer than three responses are observed in the sixteen patients; otherwise continue to stage 3. At stage 3, 26 more patients are treated. We conclude the test treatment is effective if there are more than 11 responses in the 42 patients.

11.10 COMPARING VARIABILITIES

In clinical research, when comparing a test treatment and a control or a reference treatment, the treatment effect is usually established by comparing mean response change from the baseline of some predetermined study endpoints between treatment groups, assuming that their corresponding variabilities are comparable. In practice, however, the variabilities associated with the test treatment and the control or reference treatment could be very different. When the variability of the test treatment is much larger than that of the control or reference treatment, the safety of the test treatment could be a concern. Thus, in addition to comparing mean responses between treatment groups, it is also of interest to compare the variabilities associated with the responses between treatment groups. In general, variabilities can be classified into two categories, namely, the intrasubject (or within-subject) variability and the intersubject (or between-subject) variability. Intrasubject variability refers to the variability observed from repeated measurements from the same subject under the same experimental condition, whereas intersubject variability is the variability due to the heterogeneity among subjects. The total variability is simply the sum of the intrasubject and intersubject variabilities.

For clinical trials intended for comparing variabilities, sample size calculation may be performed based on appropriate statistical methods for testing equality, noninferiority/superiority, and similarity (equivalence) between treatment groups. The problem of comparing intrasubject variabilities is well studied by Chinchilli and Esinhart (1996) through an F statistic under a replicated crossover model. A similar idea can also be applied to comparing total variabilities under a parallel design without replicates. For comparing intersubject and total variabilities under a crossover design, statistical methods developed by Lee et al. (2003) and Lee et al. (2002) are useful (see also Chow, 2003; Chow et al., 2003b).

In what follows, statistical tests for equality, noninferiority/superiority, and similarity (equivalence) in intrasubject variability and intersubject variability under a parallel-group

design with replicates or a replicated crossover design are briefly described. Also included are corresponding formulas for sample size calculation.

11.10.1 Comparing Intrasubject Variabilities

Parallel Design with Replicates Let x_{ijk} be the observation of the k th replicate ($k = 1, \dots, m$) of the j th subject ($j = 1, \dots, n_i$) from the i th treatment ($i = T, R$). It is assumed that

$$x_{ijk} = \mu_i + S_{ij} + e_{ijk}, \quad (11.10.1)$$

where μ_i is the treatment effect, S_{ij} is the random effect due to the j th subject in the i th treatment group, and e_{ijk} is the intrasubject variability under the i th treatment. It is assumed that for a fixed i , S_{ij} are independent and identically distributed as normal random variables with mean 0 and variance σ_{Bi}^2 , and e_{ijk} , $k = 1, \dots, m$ are independent and identically distributed as a normal random variable with mean 0 and variance σ_{Wi}^2 . Under this model, an unbiased estimator for σ_{Wi}^2 is given by

$$\hat{\sigma}_{Wi}^2 = \frac{1}{n_i(m-1)} \sum_{j=1}^{n_i} \sum_{k=1}^m (x_{ijk} - \bar{x}_{ij.})^2, \quad (11.10.2)$$

where

$$\bar{x}_{ij.} = \frac{1}{m} \sum_{k=1}^m x_{ijk}. \quad (11.10.3)$$

It can be seen that $n_i(m-1)\hat{\sigma}_{Wi}^2/\sigma_{Wi}^2$ is distributed as a $\chi_{n_i(m-1)}^2$ random variable.

Test for Equality For testing equality in intrasubject variability between treatment groups, the following hypotheses are often considered:

$$H_0: \sigma_{WT} = \sigma_{WR} \quad \text{vs.} \quad H_a: \sigma_{WT} \neq \sigma_{WR}.$$

A commonly used test statistic for testing the above hypotheses is given by

$$T = \frac{\hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2}.$$

Under the null hypothesis, T is distributed as an F random variable with $n_T(m-1)$ and $n_R(m-1)$ degrees of freedom. Hence, we reject the null hypothesis at the α level of significance if

$$T > F(\alpha/2, n_T(m-1), n_R(m-1))$$

or

$$T < F(1-\alpha/2, n_T(m-1), n_R(m-1)),$$

where $F(\alpha/2, n_T(m-1), n_R(m-1))$ is the upper $(\alpha/2)$ th quantile of an F distribution with $n_T(m-1)$ and $n_R(m-1)$ degrees of freedom. Under the alternative hypothesis, without

loss of generality, we assume that $\sigma_{WT}^2 < \sigma_{WR}^2$. The power of the above test can then be approximated by

$$\begin{aligned} 1 - \beta &= P(T < F(1 - \alpha/2, n_T(m-1), n_R(m-1))) \\ &= P(1/T > F(\alpha/2, n_R(m-1), n_T(m-1))) \\ &= P\left(\frac{\hat{\sigma}_{WR}^2/\sigma_{WR}^2}{\hat{\sigma}_{WT}^2/\sigma_{WT}^2} > \frac{\sigma_{WT}^2}{\sigma_{WR}^2} F(\alpha/2, n_R(m-1), n_T(m-1))\right) \\ &= P\left(F(n_R(m-1), n_T(m-1)) > \frac{\sigma_{WT}^2}{\sigma_{WR}^2} F(\alpha/2, n_R(m-1), n_T(m-1))\right), \end{aligned}$$

where $F(a, b)$ denotes an F random variable with a and b degrees of freedom. Under the assumption that $n = n_R = n_T$ and with a fixed σ_{WT}^2 and σ_{WR}^2 , the sample size needed in order to achieve a desired power of $1 - \beta$ can be obtained by solving the following equation for n :

$$\frac{\sigma_{WT}^2}{\sigma_{WR}^2} = \frac{F(1 - \beta, n(m-1), n(m-1))}{F(\alpha/2, n(m-1), n(m-1))}.$$

Test for Noninferiority/Superiority As indicated in Chow et al. (2003), the problem of testing noninferiority and superiority can be unified by the following hypotheses:

$$H_0: \frac{\sigma_{WT}}{\sigma_{WR}} \geq \delta \quad \text{vs.} \quad H_a: \frac{\sigma_{WT}}{\sigma_{WR}} < \delta$$

When $\delta < 1$, the rejection of the null hypothesis indicates the superiority of the test product over the reference in terms of the intrasubject variability. When $\delta > 1$, the rejection of the null hypothesis indicated the noninferiority of the test product over the reference. Consider the following test statistic:

$$T = \frac{\hat{\sigma}_{WT}^2}{\delta^2 \hat{\sigma}_{WR}^2}.$$

Under the null hypothesis, T is distributed as an F random variable with $n_T(m-1)$ and $n_R(m-1)$ degrees of freedom. Hence, we reject the null hypothesis at the α level of significance if

$$T < F(1 - \alpha, n_T(m-1), n_R(m-1)).$$

Under the alternative hypothesis that $\sigma_{WT}^2/\sigma_{WR}^2 < \delta$, the power of the above test can be approximated by

$$\begin{aligned} 1 - \beta &= P(T < F(1 - \alpha, n_T(m-1), n_R(m-1))) \\ &= P(1/T > F(\alpha, n_R(m-1), n_T(m-1))) \\ &= P\left(\frac{\hat{\sigma}_{WR}^2/\sigma_{WR}^2}{\hat{\sigma}_{WT}^2/\sigma_{WT}^2} > \frac{\sigma_{WT}^2}{\delta \sigma_{WR}^2} F(\alpha, n_R(m-1), n_T(m-1))\right) \\ &= P\left(F(n_R(m-1), n_T(m-1)) > \frac{\sigma_{WT}^2}{\delta^2 \sigma_{WR}^2} F(\alpha, n_R(m-1), n_T(m-1))\right). \end{aligned}$$

Under the assumption that $n = n_T = n_R$, the sample size needed in order to achieve a desired power of $1 - \beta$ at the α level of significance can be obtained by solving the following equation for n :

$$\frac{\sigma_{WT}^2}{\delta^2 \sigma_{WR}^2} = \frac{F(1-\beta, n(m-1), n(m-1))}{F(\alpha, n(m-1), n(m-1))}.$$

Test for Similarity or Equivalence For testing similarity or equivalence in intrasubject variability between treatment groups, consider the following hypotheses:

$$H_0: \frac{\sigma_{WT}^2}{\sigma_{WR}^2} \geq \delta \text{ or } \frac{\sigma_{WT}^2}{\sigma_{WR}^2} \leq 1/\delta \quad \text{vs.} \quad H_a: \frac{1}{\delta} < \frac{\sigma_{WT}^2}{\sigma_{WR}^2} < \delta,$$

where $\delta < 1$ is the similarity limit. The above hypotheses can be decomposed into the following two one-sided hypotheses:

$$H_{01}: \frac{\sigma_{WT}^2}{\sigma_{WR}^2} \geq \delta \quad \text{vs.} \quad H_{a1}: \frac{\sigma_{WT}^2}{\sigma_{WR}^2} < \delta$$

and

$$H_{02}: \frac{\sigma_{WT}^2}{\sigma_{WR}^2} \leq \frac{1}{\delta} \quad \text{vs.} \quad H_{a2}: \frac{\sigma_{WT}^2}{\sigma_{WR}^2} > \frac{1}{\delta}.$$

The following two test statistics are useful in testing the above two one-sided hypotheses:

$$T_1 = \frac{\hat{\sigma}_{WT}}{\delta \hat{\sigma}_{WR}} \quad \text{and} \quad T_2 = \frac{\delta \hat{\sigma}_{WT}}{\hat{\sigma}_{WR}}.$$

We then reject the null hypothesis and conclude similarity at the α level of significance if

$$T_1 < F(1-\alpha, n_T(m-1), n_R(m-1)) \quad \text{and} \quad T_2 > F(\alpha, n_T(m-1), n_R(m-1)),$$

Assuming that $n = n_T = n_R$ and $\sigma_{WT} = \sigma_{WR}$, under the alternative hypothesis that the power of the above test can be approximated by

$$\begin{aligned} 1 - \beta &= P \left(\frac{F(\alpha, n(m-1), n(m-1))}{\delta^2} < \frac{\hat{\sigma}_{WT}}{\hat{\sigma}_{WR}} < \delta^2 F(1-\alpha, n(m-1), n(m-1)) \right) \\ &= P \left(\frac{1}{F(1-\alpha, n(m-1), n(m-1)) \delta^2} < \frac{\hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2} < \delta^2 F(1-\alpha, n(m-1), n(m-1)) \right) \\ &\geq 1 - 2P \left(\frac{\hat{\sigma}_{WR}^2}{\hat{\sigma}_{WT}^2} > \delta^2 F(1-\alpha, n(m-1), n(m-1)) \right) \\ &= 1 - 2P \left(F(n(m-1), n(m-1)) > \frac{\delta^2 \sigma_{WT}^2}{\sigma_{WR}^2} F(1-\alpha, n(m-1), n(m-1)) \right). \end{aligned}$$

Thus, a conservative estimate for the sample size required for achieving a desired power of $1 - \beta$ can be obtained by solving the following equation for n :

$$\frac{\delta^2 \sigma_{WT}^2}{\sigma_{WR}^2} = \frac{F(\beta/2, n(m-1), n(m-1))}{F(1-\alpha, n(m-1), n(m-1))}.$$

Example 11.10.1 Suppose an investigator is interested in conducting a clinical trial under a two-arm parallel trial with 3 ($m = 3$) replicates per subject for comparing the variability of an inhaled formulation (treatment) with a subcutaneous (SC) injected formulation (control) of a compound. The primary study endpoint is area under the blood concentration time curve (AUC). Based on pharmacokinetic data observed from a pilot study, it is assumed the true variance of treatment and control are given by 30% ($\sigma_{WT}^2 = 0.30$) and 45% ($\sigma_{WR}^2 = 0.45$), respectively.

For testing equality in intrasubject variability, the sample size needed in order to achieve an 80% power at the 5% level of significance can be obtained by solving the following equation:

$$\frac{0.30}{0.45} = \frac{F(0.80, 2n, 2n)}{F(0.025, 2n, 2n)}.$$

This gives $n = 96$. The sample size required for testing noninferiority/superiority and/or similarity/equivalence can be similar using the formulas given above.

Replicated Crossover Design Unlike parallel-group designs with replicates, one of the advantages of a replicated crossover design is that it allows comparisons within subjects. For convenience's sake, we consider a $2 \times 2m$ replicated crossover design comparing two treatments (a test treatment and a control or reference treatment). Under a $2 \times 2m$ replicated crossover design, in each sequence, each subject receives the test treatment m times and the control treatment m times at different dosing periods. When $m = 1$, the $2 \times 2m$ replicated crossover design reduces to the standard two-sequence, two-period (2×2) crossover design. On the other hand, when $m = 2$, the $2 \times 2m$ replicated crossover design becomes a 2×4 crossover design, which is a design that is recommended by the FDA for assessment of population/individual bioequivalence (FDA, 2001).

Suppose that n_1 subjects are assigned to the first sequence and n_2 subjects are assigned to the second sequence. Let x_{ijkl} be the observation from the j th subject ($j = 1, \dots, n_i$) in the i th sequence ($i = 1, 2$) under the l th replicate ($l = 1, \dots, m$) of the k th treatment ($k = T, R$). Chinchilli and Esinhart (1996) proposed the following mixed effects model:

$$x_{ijkl} = \mu_k + \gamma_{ikl} + S_{ijk} + \varepsilon_{ijkl}, \quad (11.10.4)$$

where μ_k is the treatment effect for treatment k , γ_{ikl} is the fixed effect of the l th replicate on treatment k in the i th sequence with constraint

$$\sum_{i=1}^2 \sum_{l=1}^m \gamma_{ikl} = 0.$$

$(S_{ijT}, S_{ijR})'$'s are the random effects of the j th subject in the i th sequence, which are independent and identically distributed as a bivariate normal random vector with mean $(0, 0)'$ and covariance matrix

$$\Sigma_B = \begin{pmatrix} \sigma_{BT}^2 & \rho\sigma_{BT}\sigma_{BR} \\ \rho\sigma_{BT}\sigma_{BR} & \sigma_{BR}^2 \end{pmatrix}.$$

Note that σ_{BT}^2 and σ_{BR}^2 are intersubject variances under the test treatment and the control or reference treatment, respectively. ε_{ijkl} 's are independent random variables from the normal distribution with mean 0 and variance σ_{WT}^2 or σ_{WR}^2 , which are intrasubject variabilities under the test drug and the reference drug, respectively. It is assumed that $(S_{ijT}, S_{ijR})'$ and ε_{ijkl} are independent.

To obtain estimators of intrasubject variances, it is a common practice to use an orthogonal transformation, which is considered by Chinchilli and Esinhart (1996). A new random variable z_{ijkl} can be obtained by using the orthogonal transformation

$$\mathbf{z}_{ijk} = \mathbf{P}' \mathbf{x}_{ijk}, \quad (11.10.5)$$

where

$$\mathbf{x}'_{ijk} = (x_{ijk1}, x_{ijk2}, \dots, x_{ijkm}), \quad \mathbf{z}'_{ijk} = (z_{ijk1}, z_{ijk2}, \dots, z_{ijkm})$$

and \mathbf{P} is an $m \times m$ orthogonal transformation under which $\mathbf{P}'\mathbf{P}$ is a $m \times m$ diagonal matrix. The first column of \mathbf{P} is usually defined by the vector $\frac{1}{m} (1, 1, \dots, 1)'$ to obtain $z_{ijk1} = \bar{x}_{ijk}$, and the other columns can be defined to satisfy the orthogonality of \mathbf{P} and $Var(z_{ijkl}) = \sigma_{WT}^2$ for $l = 2, \dots, m$. For example, in a 2×4 crossover design, the new random variable z_{ijkl} can be defined as

$$z_{ijk1} = \frac{x_{ijk1} + x_{ijk2}}{2} = \bar{x}_{ijk} \quad \text{and} \quad z_{ijk2} = \frac{x_{ijk1} - x_{ijk2}}{\sqrt{2}}.$$

Now, the estimator of intrasubject variance can be defined as

$$\hat{\sigma}_{WT}^2 = \frac{1}{(n_1 + n_2 - 2)(m - 1)} \sum_{i=1}^2 \sum_{j=1}^{n_i} \sum_{l=2}^m (z_{ijl} - \bar{z}_{i.l})^2,$$

$$\hat{\sigma}_{WR}^2 = \frac{1}{(n_1 + n_2 - 2)(m - 1)} \sum_{i=1}^2 \sum_{j=1}^{n_i} \sum_{l=2}^m (z_{ijl} - \bar{z}_{i.Rl})^2,$$

where

$$\bar{z}_{i.kl} = \frac{1}{n_i} \sum_{j=1}^{n_i} z_{ijl}.$$

It should be noted that $\hat{\sigma}_{WT}^2$ and $\hat{\sigma}_{WR}^2$ are independent and distributed as F -distribution, respectively.

Test for Equality Similarly, we consider the following hypotheses:

$$H_0 : \sigma_{WT} = \sigma_{WR} \quad \text{vs.} \quad H_a : \sigma_{WT} \neq \sigma_{WR}$$

for testing equality in intrasubject variability.

Under the null hypothesis, test statistic

$$T = \frac{\hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2}$$

is distributed as an F random variable with d and d degrees of freedom, where $d = (n_1 + n_2 - 2)(m - 1)$. Hence, we reject the null hypothesis at the α level of significance if

$$T > F(\alpha/2, d, d)$$

or

$$T < F(1 - \alpha/2, d, d).$$

Under the alternative hypothesis, without loss of generality, we assume that $\hat{\sigma}_{WT}^2 < \hat{\sigma}_{WR}^2$. The power of the above test can then be approximated by

$$\begin{aligned} 1 - \beta &= P(T < F(1 - \alpha/2, d, d)) \\ &= P(1/T > F(\alpha/2, d, d)) \\ &= P\left(\frac{\hat{\sigma}_{WR}^2/\sigma_{WR}^2}{\hat{\sigma}_{WT}^2/\sigma_{WT}^2} > \frac{\hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2} F(\alpha/2, d, d)\right) \\ &= P\left(F(d, d) > \frac{\sigma_{WT}^2}{\sigma_{WR}^2} F(\alpha/2, d, d)\right). \end{aligned}$$

Under the assumption that $n = n_1 = n_2$ and with fixed $\hat{\sigma}_{WT}^2$ and $\hat{\sigma}_{WR}^2$, the sample size needed in order to achieve a desired power of $1 - \beta$ can be obtained by solving the following equation for n :

$$\frac{\sigma_{WT}^2}{\sigma_{WR}^2} = \frac{F(1 - \beta, (2n - 2)(m - 1), (2n - 2)(m - 1))}{F(\alpha/2, (2n - 2)(m - 1), (2n - 2)(m - 1))}.$$

Test for Noninferiority/Superiority Similarly, we consider the following hypotheses:

$$H_0 : \frac{\sigma_{WT}}{\sigma_{WR}} \geq \delta \quad \text{vs.} \quad H_a : \frac{\sigma_{WT}}{\sigma_{WR}} < \delta$$

for testing noninferiority and superiority. Consider the following test statistic:

$$T = \frac{\hat{\sigma}_{WT}^2}{\delta^2 \hat{\sigma}_{WR}^2}.$$

Under the null hypothesis, T is distributed as an F random variable with d and d degrees of freedom. Hence, we reject the null hypothesis at the α level of significance if

$$T < F(1 - \alpha, d, d).$$

Under the alternative hypothesis that $\sigma_{WT}^2/\sigma_{WR}^2 < \delta$, the power of the above test can be approximated by

$$\begin{aligned} 1 - \beta &= P(T < F(1 - \alpha, d, d)) \\ &= P(1/T > F(\alpha, d, d)) \\ &= P\left(\frac{\hat{\sigma}_{WR}^2/\sigma_{WR}^2}{\hat{\sigma}_{WT}^2/\sigma_{WT}^2} > \frac{\sigma_{WT}^2}{\delta^2 \sigma_{WR}^2} F(\alpha, d, d)\right) \\ &= P\left(F(d, d) > \frac{\sigma_{WT}^2}{\delta^2 \sigma_{WR}^2} F(\alpha, d, d)\right). \end{aligned}$$

Thus, under the assumption that $n = n_1 = n_2$, the sample size needed in order to achieve a desired power of $1 - \beta$ at the α level of significance can be obtained by solving the following equation for n :

$$\frac{\sigma_{WT}^2}{\delta^2 \sigma_{WR}^2} = \frac{F(1 - \beta, (2n - 2)(m - 1), (2n - 2)(m - 1))}{F(\alpha, (2n - 2)(m - 1), (2n - 2)(m - 1))}.$$

Test for Similarity or Equivalence For testing similarity or equivalence in intrasubject variability, similarly, we consider the following two one-sided hypotheses:

$$H_{01} : \frac{\sigma_{WT}}{\sigma_{WR}} \geq \delta \quad \text{vs.} \quad H_{a1} : \frac{\sigma_{WT}}{\sigma_{WR}} < \delta$$

and

$$H_{02} : \frac{\sigma_{WT}}{\sigma_{WR}} \leq \frac{1}{\delta} \quad \text{vs.} \quad H_{a2} : \frac{\sigma_{WT}}{\sigma_{WR}} > \frac{1}{\delta}.$$

These two hypotheses can be tested by the following two test statistics:

$$T_1 = \frac{\hat{\sigma}_{WT}^2}{\delta^2 \hat{\sigma}_{WR}^2} \quad \text{and} \quad T_2 = \frac{\delta^2 \hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2}.$$

We then reject the null hypothesis and conclude similarity at the α level of significance if

$$T_1 < F(1 - \alpha, d, d) \quad \text{and} \quad T_2 > F(\alpha, d, d).$$

Assuming that $n = n_1 = n_2$, under the alternative hypothesis that $\hat{\sigma}_{WT}^2 \leq \hat{\sigma}_{WR}^2$, the power of the above test can be approximated by

$$\begin{aligned} 1 - \beta &= P\left(\frac{F(\alpha, d, d)}{\delta} < \frac{\hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2} < \delta F(1 - \alpha, d, d)\right) \\ &= P\left(\frac{1}{F(1 - \alpha, d, d) \delta} < \frac{\hat{\sigma}_{WT}^2}{\hat{\sigma}_{WR}^2} < \delta F(1 - \alpha, d, d)\right) \\ &\geq 1 - 2P\left(\frac{\hat{\sigma}_{WR}^2}{\hat{\sigma}_{WT}^2} > \delta^2 F(1 - \alpha, d, d)\right) \\ &= 1 - 2P\left(F(d, d) > \frac{\delta^2 \sigma_{WT}^2}{\sigma_{WR}^2} F(1 - \alpha, d, d)\right). \end{aligned}$$

Hence, a conservative estimate for the sample size needed in order to achieve the power of $1 - \beta$ can be obtained by solving the following equation:

$$\frac{\delta^2 \sigma_{WT}^2}{\sigma_{WR}^2} = \frac{F(\beta/2, (2n - 2)(m - 1), (2n - 2)(m - 1))}{F(1 - \alpha, (2n - 2)(m - 1), (2n - 2)(m - 1))}.$$

Example 11.10.2 Consider Example 11.10.1 described above. Suppose the intended study will be conducted under a 2×4 replicated crossover design rather than a parallel design with three replicates.

It is assumed that the true variance of inhaled formulation and SC formulation are given by 30% ($\sigma_{WT}^2 = 0.30$) and 45% ($\sigma_{WR}^2 = 0.45$), respectively. Hence, the sample size needed for testing equality in intrasubject variability for achieving an 80% power at the 5% level of significance can be obtained by solving the following equation:

$$\frac{0.30}{0.45} = \frac{F(0.80, 2n - 2, 2n - 2)}{F(0.025, 2n - 2, 2n - 2)}.$$

This gives $n = 98$.

Under the 2×4 replicated crossover design, the sample sizes required for testing non-inferiority, superiority, and similarity or equivalence can be similarly obtained using the formulas given above.

11.10.2 Comparing Intersubject Variabilities

In clinical research, it is not uncommon that clinical results may not be reproducible from subject to subject within the target population or from subjects within the target population to subjects within a similar but slightly different population due to the intersubject variability. As pointed out in Chow et al. (2003b), the following difficulties are usually encountered when testing a difference in intersubject variabilities. First, unbiased estimators of the intersubject and total variabilities are usually not chi-square distributed under both parallel and crossover designs with replicates. Second, the estimators for the intersubject and total variabilities under different treatments are usually not independent under a crossover design. As a result, unlike tests for comparing intrasubject variabilities, the standard F test is not applicable. Tests for comparing intersubject variabilities under a parallel design can be performed by using the method of a modified large sample (MLS) method. See, e.g., Howe (1974), Graybill and Wang (1980), Ting et al. (1990), Hyslop et al. (2000), and Lee et al. (2003).

Parallel Design with Replicates Under model (11.10.1), define

$$s_{Bi}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\bar{x}_{ij..} - \bar{x}_{i..})^2, \quad (11.10.6)$$

where $\bar{x}_{ij..}$ is given in (11.10.3) and

$$\bar{x}_{i..} = \frac{1}{n_i} \sum_{j=1}^{n_i} \bar{x}_{ij..}$$

Note that $E(s_{Bi}^2) = \sigma_{Bi}^2 + \sigma_{Wi}^2/m$. Therefore, the unbiased estimators for the intersubject variances are given by

$$\hat{\sigma}_{Bi}^2 = s_{Bi}^2 - \frac{1}{m} \hat{\sigma}_{Wi}^2,$$

where $\hat{\sigma}_{Wi}^2$ is defined in (11.10.2).

Test for Equality For testing equality in intersubject variability, the following hypotheses are usually considered:

$$H_0: \frac{\sigma_{BT}}{\sigma_{BR}} = 1 \quad \text{vs.} \quad H_a: \frac{\sigma_{BT}}{\sigma_{BR}} \neq 1.$$

Testing the above hypotheses is equivalent to testing the following hypotheses:

$$H_0: \sigma_{BT}^2 - \sigma_{BR}^2 = 0 \quad \text{vs.} \quad H_a: \sigma_{BT}^2 - \sigma_{BR}^2 \neq 0.$$

Let $\eta = \sigma_{BT}^2 - \sigma_{BR}^2$. An intuitive estimator of η is given by

$$\hat{\eta} = \hat{\sigma}_{BT}^2 - \hat{\sigma}_{BR}^2.$$

It follows that

$$\begin{aligned}\hat{\eta} &= \hat{\sigma}_{BR}^2 - \hat{\sigma}_{BT}^2 \\ &= s_{BT}^2 - s_{BR}^2 - \hat{\sigma}_{WT}^2/m + \hat{\sigma}_{WR}^2/m.\end{aligned}$$

An $(1-\alpha) \times 100\%$ confidence interval of η is given by

$$(\hat{\eta}_L, \hat{\eta}_U),$$

where

$$\hat{\eta}_L = \hat{\eta} - \sqrt{\Delta_L}, \quad \hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U}$$

and

$$\begin{aligned}\Delta_L &= s_{BT}^4 \left(1 - \frac{n_T - 1}{\chi^2(\alpha/2, n_T - 1)}\right)^2 + s_{BR}^4 \left(1 - \frac{1 - n_R - 1}{\chi^2(1 - \alpha/2, n_R - 1)}\right)^2 \\ &\quad + \frac{\hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_T(m - 1)}{\chi^2(1 - \alpha/2, n_T(m - 1))}\right)^2 + \frac{\hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{1 - n_R(m - 1)}{\chi^2(\alpha/2, n_R(m - 1))}\right)^2 \\ \Delta_U &= s_{BT}^4 \left(1 - \frac{n_T - 1}{\chi^2(1 - \alpha/2, n_T - 1)}\right)^2 + s_{BR}^4 \left(1 - \frac{n_R - 1}{\chi^2(\alpha/2, n_R - 1)}\right)^2 \\ &\quad + \frac{\hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_T(m - 1)}{\chi^2(\alpha/2, n_T(m - 1))}\right)^2 + \frac{\hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_R(m - 1)}{\chi^2(1 - \alpha/2, n_R(m - 1))}\right)^2.\end{aligned}$$

We reject the null hypothesis at the α level of significance if $0 \notin (\hat{\eta}_L, \hat{\eta}_U)$. Under the alternative hypothesis, without loss of generality, we assume that $\sigma_{BR}^2 > \sigma_{BT}^2$ and $n = n_T = n_R$. Thus, the power of the above test procedure can be approximated by

$$P\left\{N\left(\frac{\sqrt{n}(\sigma_{BT}^2 - \sigma_{BR}^2)}{\sigma^*}, 1\right) > z(\alpha/2)\right\},$$

where

$$\begin{aligned}\sigma^{*2} &= 2 \left[\left(\sigma_{BT}^2 + \frac{\sigma_{WT}^2}{m} \right)^2 + \left(\sigma_{BR}^2 + \frac{\sigma_{WR}^2}{m} \right)^2 \right. \\ &\quad \left. + \frac{\sigma_{WT}^4}{m^2(m - 1)} + \frac{\sigma_{WR}^4}{m^2(m - 1)} \right].\end{aligned}$$

As a result, the sample size needed in order to achieve the desired power of $1 - \beta$ at the α level of significance can be obtained by solving the following equation:

$$Z(\alpha/2) - \frac{\sqrt{n}(\sigma_{BT}^2 - \sigma_{BR}^2)}{\sigma^*} = Z(\beta).$$

This leads to

$$n = \frac{\sigma^{*2}(Z(\alpha/2) + Z(\beta))^2}{(\sigma_{BT}^2 - \sigma_{BR}^2)^2}.$$

Test for Noninferiority/Superiority Similar to testing intrasubject variabilities, the problem of testing noninferiority/superiority can be unified by the following hypotheses:

$$H_0: \frac{\sigma_{BT}}{\sigma_{BR}} \geq \delta \quad \text{vs.} \quad H_a: \frac{\sigma_{BT}}{\sigma_{BR}} < \delta.$$

Testing the above hypotheses is equivalent to testing the following hypotheses:

$$H_0: \sigma_{BT}^2 - \delta^2 \sigma_{BR}^2 \geq 0 \quad \text{vs.} \quad H_1: \sigma_{BT}^2 - \delta^2 \sigma_{BR}^2 < 0.$$

Define

$$\eta = \sigma_{BT}^2 - \delta^2 \sigma_{BR}^2.$$

For a given significance level α , similarly, the $(1 - \alpha) \times 100\%$ th MLS upper confidence bound of η can be constructed as

$$\hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U},$$

where Δ_U is given by

$$\begin{aligned} \Delta_U = & s_{BT}^4 \left(1 - \frac{n_T - 1}{\chi^2(\alpha, n_T - 1)} \right)^2 + \delta^4 s_{BR}^4 \left(1 - \frac{n_R - 1}{\chi^2(\alpha, n_R - 1)} \right)^2 \\ & + \frac{\hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_T(m - 1)}{\chi^2(\alpha, n_T(m - 1))} \right)^2 + \frac{\delta^4 \hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_R(m - 1)}{\chi^2(\alpha, n_R(m - 1))} \right)^2. \end{aligned}$$

We then reject the null hypothesis at the α level of significance if $\hat{\eta}_U < 0$. Under the assumptions that $n = n_T = n_R$, using a similar argument as the previous section, the power of the above testing procedure can be approximated by

$$P \left\{ N \left(\frac{\sqrt{n}(\sigma_{BT}^2 - \delta^2 \sigma_{BR}^2)}{\sigma^*}, 1 \right) > Z(\alpha) \right\},$$

where

$$\begin{aligned} \sigma^{*2} = & 2 \left[\left(\sigma_{BT}^2 + \frac{\sigma_{WT}^2}{m} \right)^2 + \delta^4 \left(\sigma_{BR}^2 + \frac{\sigma_{WR}^2}{m} \right)^2 \right. \\ & \left. + \frac{\sigma_{WT}^4}{m^2(m - 1)} + \frac{\delta^4 \sigma_{WR}^4}{m^2(m - 1)} \right]. \end{aligned}$$

As a result, the sample size needed in order to achieve the power of $1 - \beta$ at the α level of significance can be obtained by solving the following equation:

$$Z(\alpha) - \frac{\sqrt{n}(\sigma_{BT}^2 - \delta^2 \sigma_{BR}^2)^2}{\sigma^*} = -Z(\beta).$$

This gives

$$n = \frac{\sigma^{*2}(Z(\alpha) + Z(\beta))^2}{(\sigma_{BT}^2 - \delta^2 \sigma_{BR}^2)^2}.$$

Example 11.10.3 Consider Example 11.10.1 as described above. Suppose the investigator is interested in testing the difference in intersubject variability between inhaled and SC formulations. Suppose based on pharmacokinetic data observed from a pilot study, the intrasubject and intersubject variabilities for the treatment and the control are estimated as 20% and 30% (i.e., $\sigma_{WT} = 0.20$ and $\sigma_{WR} = 0.30$) and 35% and 40% (i.e., $\sigma_{BT} = 0.30$ and $\sigma_{BR} = 0.40$), respectively. Thus, we have

$$\begin{aligned}\sigma_{BT} &= 0.30 & \sigma_{BR} &= 0.40, \\ \sigma_{WT} &= 0.20 & \sigma_{WR} &= 0.30.\end{aligned}$$

It follows that

$$\begin{aligned}\sigma^{\ast 2} &= 2 \left[\left(0.30^2 + \frac{0.20^2}{3} \right)^2 + \left(0.40^2 + \frac{0.30^2}{3} \right)^2 \right. \\ &\quad \left. + \frac{0.20^4}{3^2(3-1)} + \frac{0.30^4}{3^2(3-1)} \right] \\ &= 0.095.\end{aligned}$$

Hence, the sample size needed for testing equality in intersubject variability for achieving an 80% power ($1 - \beta = 0.80$) at the 5% level of significance ($\alpha = 0.05$) is given by

$$\begin{aligned}n &= \frac{0.095(1.96 + 0.84)^2}{(0.30^2 - 0.40^2)^2} \\ &= 151.4 \approx 152.\end{aligned}$$

Sample sizes required for testing noninferiority, superiority, and similarity/equivalence can be similarly obtained by using the formulas given above.

Replicated Crossover Design Under model (11.10.4) and new random variable $z_{ijk1} \equiv \bar{x}_{ijk}$, in (11.10.5), the estimators of intersubject variances can be defined by

$$\begin{aligned}s_{BT}^2 &= \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^2 \sum_{j=1}^{n_i} (\bar{x}_{ijT} - \bar{x}_{iT.})^2, \\ s_{BR}^2 &= \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^2 \sum_{j=1}^{n_i} (\bar{x}_{ijR} - \bar{x}_{iR.})^2,\end{aligned}$$

where

$$\bar{x}_{i.k.} = \frac{1}{n_i} \sum_{j=1}^{n_i} \bar{x}_{ijk.}$$

Note that $E(s_{Bk}^2) = \sigma_{Bk}^2 + \sigma_{Wk}^2/m$ for $k = T, R$. Therefore, the unbiased estimators for the intersubject variance are given by

$$\hat{\sigma}_{BT}^2 = s_{BT}^2 - \frac{1}{m} \hat{\sigma}_{WT}^2, \tag{11.10.7}$$

$$\hat{\sigma}_{BR}^2 = s_{BR}^2 - \frac{1}{m} \hat{\sigma}_{WR}^2. \tag{11.10.8}$$

Test for Equality For testing the equality in intersubject variability, the following hypotheses are considered:

$$H_0: \frac{\sigma_{BT}}{\sigma_{BR}} = 1 \quad \text{vs.} \quad H_a: \frac{\sigma_{BT}}{\sigma_{BR}} \neq 1.$$

Testing the above hypotheses is equivalent to test the following hypotheses:

$$H_0: \sigma_{BT}^2 - \sigma_{BR}^2 = 0 \quad \text{vs.} \quad H_a: \sigma_{BT}^2 - \sigma_{BR}^2 \neq 0.$$

Let $\eta = \sigma_{BT}^2 - \sigma_{BR}^2$. An intuitive estimator of η is given by

$$\hat{\eta} = \hat{\sigma}_{BT}^2 - \hat{\sigma}_{BR}^2,$$

where $\hat{\sigma}_{BT}^2$ and $\hat{\sigma}_{BR}^2$ are given in (11.10.7) and (11.10.8), respectively. It follows that

$$\begin{aligned} \hat{\eta} &= \hat{\sigma}_{BR}^2 - \hat{\sigma}_{BT}^2 \\ &= s_{BT}^2 - s_{BR}^2 - \frac{1}{m}\hat{\sigma}_{WT}^2 + \frac{1}{m}\hat{\sigma}_{WR}^2. \end{aligned}$$

Random vector $(\bar{x}_{ijT}, \bar{x}_{ijR})'$ for the j th subject in i th sequence has a bivariate normal distribution with covariance matrix given by

$$\Omega_B = \begin{pmatrix} \sigma_{BT}^2 + \sigma_{WT}^2/m & \rho\sigma_{BT}\sigma_{BR} \\ \rho\sigma_{BT}\sigma_{BR} & \sigma_{BR}^2 + \sigma_{WR}^2/m \end{pmatrix}. \quad (11.10.9)$$

The unbiased estimator of covariance matrix Ω_B can be obtained as

$$\hat{\Omega}_B = \begin{pmatrix} s_{BT}^2 & s_{BTR}^2 \\ s_{BTR}^2 & s_{BR}^2 \end{pmatrix}, \quad (11.10.10)$$

where

$$s_{BTR}^2 = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^2 \sum_{j=1}^{n_i} (\bar{x}_{ijT} - \bar{x}_{iT})(\bar{x}_{ijR} - \bar{x}_{iR})$$

is the sample covariance between \bar{x}_{ijT} and \bar{x}_{ijR} . Let λ_i , $i = 1, 2$ be the two eigenvalues of the matrix $\Theta\Omega_B$, where

$$\Theta = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (11.10.11)$$

Hence, λ_i , $i = 1, 2$ can be estimated by

$$\lambda_i = \frac{s_{BT}^2 - s_{BR}^2 \pm \sqrt{(s_{BT}^2 + s_{BR}^2)^2 - 4s_{BTR}^4}}{2} \quad \text{for } i = 1, 2.$$

Without loss of generality, it can be assumed that $\hat{\lambda}_1 < 0 < \hat{\lambda}_2$. By Lee et al. (2003), an $(1 - \alpha) \times 100\%$ confidence interval of η is given by

$$(\hat{\eta}_L, \hat{\eta}_U),$$

where

$$\hat{\eta}_L = \hat{\eta} - \sqrt{\Delta_L}, \quad \hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U}$$

and

$$\begin{aligned}\Delta_L &= \hat{\lambda}_1^2 \left(1 - \frac{n_s - 1}{\chi^2(\alpha/2, n_s - 1)} \right)^2 + \hat{\lambda}_2^2 \left(1 - \frac{n_s - 1}{\chi^2(1 - \alpha/2, n_s - 1)} \right)^2 \\ &\quad + \frac{\hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_s(m - 1)}{\chi^2(\alpha/2, n_s(m - 1))} \right)^2 + \frac{\hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_s(m - 1)}{\chi^2(1 - \alpha/2, n_s(m - 1))} \right)^2, \\ \Delta_U &= \hat{\lambda}_1^2 \left(1 - \frac{n_s - 1}{\chi^2(1 - \alpha/2, n_s - 1)} \right)^2 + \hat{\lambda}_2^2 \left(1 - \frac{n_s - 1}{\chi^2(\alpha/2, n_s - 1)} \right)^2 \\ &\quad + \frac{\hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_s(m - 1)}{\chi^2(1 - \alpha/2, n_s(m - 1))} \right)^2 + \frac{\hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_s(m - 1)}{\chi^2(\alpha/2, n_s(m - 1))} \right)^2,\end{aligned}$$

where $n_s = n_1 + n_2 - 2$. Then, we reject the null hypothesis at the α level of significance if $0 \notin (\hat{\eta}_L, \hat{\eta}_U)$.

Let $n = n_1 + n_2 - 2$. Under the alternative hypothesis, the power of the above test can be approximated by

$$P\left\{ N\left(\frac{\sqrt{n}(\sigma_{BT}^2 - \sigma_{BR}^2)}{\sigma^*}, 1 \right) > Z(\alpha/2) \right\},$$

where

$$\begin{aligned}\sigma^{*2} &= 2 \left[\left(\sigma_{BT}^2 + \frac{\sigma_{WT}^2}{m} \right)^2 + \left(\sigma_{BR}^2 + \frac{\sigma_{WR}^2}{m} \right)^2 - 2\rho^2 \sigma_{BT}^2 \sigma_{BR}^2 \right. \\ &\quad \left. + \frac{\sigma_{WT}^4}{m^2(m - 1)} + \frac{\sigma_{WR}^4}{m^2(m - 1)} \right].\end{aligned}$$

Thus, the sample size needed in order to achieve the power of $1 - \beta$ at the α level of significance can be obtained by solving the following equation:

$$Z(\alpha/2) - \frac{\sqrt{n}(\sigma_{BT}^2 - \sigma_{BR}^2)}{\sigma^*} = -Z(\beta).$$

This leads to

$$n = \frac{\sigma^{*2}(Z(\alpha/2) + Z(\beta))^2}{(\sigma_{BT}^2 - \sigma_{BR}^2)^2}.$$

Test for Noninferiority/Superiority Similar to testing intrasubject variabilities, the problem of testing noninferiority/superiority can be unified by the following hypotheses:

$$H_0: \frac{\sigma_{BT}}{\sigma_{BR}} \geq \delta \quad \text{vs.} \quad H_a: \frac{\sigma_{BT}}{\sigma_{BR}} < \delta.$$

Testing the above hypotheses is equivalent to testing the following hypotheses:

$$H_0: \sigma_{BT}^2 - \delta^2 \sigma_{BR}^2 \geq 0 \quad \text{vs.} \quad H_a: \sigma_{BT}^2 - \delta^2 \sigma_{BR}^2 < 0.$$

When $\delta < 1$, the rejection of the null hypothesis indicates the superiority of the test drug versus the reference drug. When $\delta > 1$, the rejection of the null hypothesis indicates the noninferiority of the test drug versus the reference drug. Let $\eta = \sigma_{BT}^2 - \delta^2 \sigma_{BR}^2$. For a given significance level of α , similarly, the $(1 - \alpha)$ th upper confidence bound of η can be constructed as

$$\hat{\eta}_U = \hat{\eta} + \sqrt{\Delta_U},$$

where Δ_U is given by

$$\begin{aligned}\Delta_U = & \hat{\lambda}_1^2 \left(1 - \frac{n_s - 1}{\chi^2(1 - \alpha, n_s - 1)} \right)^2 + \hat{\lambda}_2^2 \left(1 - \frac{n_s - 1}{\chi^2(\alpha, n_s - 1)} \right)^2 \\ & + \frac{\hat{\sigma}_{WT}^4}{m^2} \left(1 - \frac{n_s(m - 1)}{\chi^2(1 - \alpha, n_s(m - 1))} \right)^2 + \frac{\delta^4 \hat{\sigma}_{WR}^4}{m^2} \left(1 - \frac{n_s(m - 1)}{\chi^2(\alpha, n_s(m - 1))} \right)^2,\end{aligned}$$

where $n_s = n_1 + n_2 - 2$ and

$$\hat{\lambda}_i = \frac{s_{BT}^2 - \delta^2 s_{BR}^2 \pm \sqrt{(s_{BT}^2 + \delta^2 s_{BR}^2)^2 - 4\delta^2 s_{BTR}^4}}{2} \quad \text{for } i = 1, 2.$$

We then reject the null hypothesis at the α level of significance if $\hat{\eta}_U < 0$.

Using a similar argument in the previous section, the power of the above test procedure can be approximated by

$$P\left(N\left(\frac{\sqrt{n}(\sigma_{BT}^2 - \delta^2 \sigma_{BR}^2)}{\sigma^*}, 1\right) > Z(\alpha)\right),$$

where

$$\begin{aligned}\sigma^{*2} = & 2 \left[\left(\sigma_{BT}^2 + \frac{\sigma_{WT}^2}{m} \right)^2 + \delta^4 \left(\sigma_{BR}^2 + \frac{\sigma_{WR}^2}{m} \right)^2 - 2\delta^2 \rho^2 \sigma_{BT}^2 \sigma_{BR}^2 \right. \\ & \left. + \frac{\sigma_{WT}^4}{m^2(m - 1)} + \frac{\delta^4 \sigma_{WR}^4}{m^2(m - 1)} \right].\end{aligned}$$

As a result, the sample size needed in order to achieve a power of $1 - \beta$ at the α level of significance can be obtained by solving the following equation:

$$Z(\alpha) - \frac{\sqrt{n}(\sigma_{BT}^2 - \delta^2 \sigma_{BR}^2)}{\sigma^*} = -Z(\beta).$$

This leads to

$$n = \frac{\sigma^{*2}(Z(\alpha) + Z(\beta))^2}{(\sigma_{BT}^2 - \delta^2 \sigma_{BR}^2)^2}.$$

Example 11.10.4 Consider Example 11.10.3. Suppose it is of interest to compare inter-subject variabilities under a 2×4 crossover design (ABAB, BABA). Information from the

pilot study indicates that $\rho = 0.75$, $\sigma_{BT} = 0.3$, $\sigma_{BR} = 0.4$, $\sigma_{WT} = 0.2$, and $\sigma_{WR} = 0.3$. It follows that

$$\begin{aligned}\sigma^*{}^2 &= 2 \left[\left(0.30^2 + \frac{0.20^2}{2} \right)^2 + \left(0.40^2 + \frac{0.30^2}{2} \right)^2 - 2(0.75)^2 (0.3)^2 (0.4)^2 \right. \\ &\quad \left. + \frac{0.20^4}{2^2} + \frac{0.30^4}{2^2} \right] \\ &= 0.08.\end{aligned}$$

Hence, the sample size needed for testing equality in intersubject variability for achieving an 80% power ($1 - \beta = 0.80$) at the 5% level of significance ($\alpha = 0.05$) is given by

$$\begin{aligned}n &= \frac{0.08(1.96 + 0.84)^2}{(0.30^2 - 0.40^2)^2} \\ &= 129.12 \approx 130.\end{aligned}$$

Because $n = n_1 + n_2 - 2$, approximately 66 subjects per sequence are required for achieving an 80% power at the 5% level of significance.

Sample sizes required for testing noninferiority, superiority, and similarity/equivalence can be similarly obtained by using the formulas given above.

11.11 DISCUSSION

For convenience sake, we have considered $n_i = n$ for all i , where i can be the i th treatment group in parallel designs or the i th sequence in crossover designs. In practice, n_i are not necessarily the same for all i . For example, when conducting a placebo control clinical trial with very ill patients or patients with severe or life-threatening diseases, it is not ethical to put too many patients in the placebo arm. Then the investigator must put fewer patients in the placebo (if the placebo arm is considered necessary to demonstrate the effectiveness and safety of the drug under investigation). A typical ratio of patient allocation for situations of this kind is 1 : 2; that is, each patient has a one-third chance of being assigned to the placebo group and a two-third chance of receiving the active drug. For different ratios of patient allocation, the sample size formulas discussed in this chapter can be directly applied with appropriate modifications of the corresponding degrees of freedom (see, e.g., Chow et al., 2003b).

It should be noted that the sample size obtained based on the formulas in this chapter is the number of *evaluable* patients required in order to achieve a desired power. In practice, we may have to enroll more patients in order to obtain the required evaluable patients due to potential dropout. Therefore, during the stage of planning, a sample size that can account for the potential dropout is usually selected. For example, if the sample size required for an intended clinical trial is n and the potential dropout rate is p , then we need to enroll $n/(1-p)^2$ patients in order to obtain n evaluable patient at the completion of the trial (Lachin, 1981). It should also be noted that the investigator may have to screen more patients in order to obtain $n/(1-p)^2$ *qualified* patients at the entry of the study based on inclusion/exclusion criteria of the trial.

In many clinical trials, multiple comparisons may be performed. In the interest of controlling the overall type I error rate at the α level an adjustment for multiple comparisons such as the Bonferroni adjustment is necessary. The formulas for sample size determination discussed in this chapter can still be applied by simply replacing the α level with an adjusted α level.

Fleiss (1986a) points out that the required sample size may be reduced if the response variable can be described by a covariate. Let n be the required sample size per group when the design does not call for the experimental control of a prognostic factor. Also let n^* be the required sample size for the study with the factor controlled. The relative efficiency (RE) between the two designs is defined as

$$RE = \frac{n}{n^*}.$$

As indicated by Fleiss (1986a), if the correlation between the prognostic factor (covariate) and the response variable is r . The RE can be expressed as

$$RE = \frac{100}{1 - r^2}.$$

Hence we have

$$n^* = n(1 - r^2).$$

As a result the required sample size per group can be reduced if the correlation exists. For example, a correlation of $r = 0.32$ could result in a 10% reduction in the sample size.

In most clinical trials, although the primary objectives are usually to evaluate the effectiveness and safety of the test drug under study, the assessment of drug safety has not received the same level of attention as the assessment of efficacy. As a result sample size calculations are usually performed based on a pre-study power analysis for the primary efficacy variable. If the sample size is based on a primary safety variable such as the adverse event rate, a large sample size may be required especially when the incidence rate is low. For example, if the incidence rate is one per 10,000, then we will need to include 10,000 in order to observe one incidence. O'Neill (1988a) indicates that the magnitude of rates that can feasibly be studied in most clinical trials is about 0.01 and higher. However, observational cohort studies usually can assess rates on the order of 0.001 and higher. O'Neill (1988a) also indicates that it is informative to examine the sample sizes that are needed to estimate a rate or to detect or estimate differences of specified amounts between the rates for two different treatment groups.

12

ISSUES IN EFFICACY EVALUATION

12.1 INTRODUCTION

As was discussed in Chapter 1, the characteristics of an adequate, well-controlled clinical trial include a study protocol with a valid statistical design, adequate controls, appropriate randomization and blinding procedures, the choice of sensitive efficacy and safety clinical endpoints, a strict adherence to the study protocol during the conduct of the trial, and a sound statistical analysis. These components are crucial for providing a scientific and unbiased assessment of the effectiveness and safety of a drug product. After the completion of a study, it is extremely important to summarize the clinical results and provide a valid scientific interpretation.

To assist the sponsors in the preparation of final clinical reports for regulatory submission and review, most regulatory agencies have developed guidelines for the format and content of a clinical report. In the United States, before the mid-1980s, clinical study reports were typically prepared by the study's medical monitor using the study's statistical reports. As a result, for each clinical study, two reports were submitted to the FDA for review. The clinical report was reviewed by the respective therapeutic division and the statistical report was reviewed by the biometrics division. The disadvantage of this arrangement is that the clinical reports may not provide the medical reviewers with the sufficient information for an adequate and correct interpretation of the results of statistical analyses. On the other hand, statistical reports often lack important clinical interpretations for the statistical reviewers to appreciate and understand the clinical significance and the magnitude of the study findings. To overcome this disadvantage, the FDA made a revolutionary change in the reporting of a clinical trial by publishing the *Guideline for the Format and Content of the Clinical and Statistical Sections of an Application* in July 1988. This guideline specifies that two separate clinical and statistical reports must be combined into a full integrated report. The full

integrated report cannot be derived by simply attaching a separate statistical report to the clinical report. The advantage of this revolutionary change is that it forces both medical and statistical reviewers to review the relevant material and information in a single report.

In 1994 the Committee for Proprietary Material Products (CPMP) Working Party on Efficacy on Medicinal Products of the European Community issued a similar guideline entitled *A Note for Guidance on Biostatistical Methodology in Clinical Trials in Applications for Marketing Authorizations for Medical Products* (CPMP, 1995). In July, 1996, the International Conference on Harmonization (ICH) issued a guideline entitled *the Structure and Contents of Clinical Study Reports*. The ICH recommended its adoption to the three regulatory agencies. Technically speaking, clinical and statistical reports must now include clinical data listing, statistical methods, and results of statistical analysis in addition to documentation on the conduct of the clinical trial. For an integrated clinical and statistical report, the FDA guidelines suggest that tables and figures be incorporated into the main text of the report or placed at the end of the text. The study protocol, investigator information, related publications, patient data listings, and technical statistical details such as derivations, computations, analyses, and computer output must be provided in the appendixes of the report.

In clinical trials the observed clinical data can basically be summarized at three different levels. The first level concerns patient data listings which include death, discontinued patients, protocol deviations, patients excluded from the efficacy analysis, listings of serious adverse events, laboratory abnormal values, individual demographic, efficacy and adverse events listings, and laboratory measurements. The FDA also requires that the case report tabulations be included at this level. The second level concerns statistical tables and figures provided at the end but not in the text of the report. These statistical tables and figures are derived from the results of the designated statistical methods stated in the protocol. The third level concerns the summary tables and figures in the text which are condensed from the summary statistical tables and figures. Cross-references must be provided to summary tables and figures in the text, statistical tables and figures at the end of the study, and individual patient data listings in the appendices of the study.

For efficacy evaluation of the drug product under investigation, formal statistical inferential procedures are usually performed to establish the benefit of the treatment based on the efficacy data. However, many clinical and statistical issues may be raised during the analysis of efficacy data. These issues need to be addressed before a fair and unbiased assessment of efficacy can be reached. The FDA guidelines on the format and content for the full integrated clinical and statistical report and the ICH guidelines on the structure and contents of clinical study reports require the following issues be addressed in the final reports: (1) baseline comparability, (2) analyses of the intention-to-treat dataset versus evaluable dataset, (3) adjustments of covariates, (4) multicenter trials, (5) subgroups analysis, (6) multiple endpoints, (7) interim analysis and data monitoring, (8) active control studies, and (9) handling of dropouts or missing data. Due to the globalization and recent advancement on genetics and bioinformatics, combining efficacy evidence from different geographic regions with different ethnic backgrounds and use of genomic information to account for variation among regions and ethnicities have become a critical yet challenging issue for evaluation of the efficacy of the medicine under investigation. Note that the issues of active control studies and equivalence trials were previously addressed in Section 7.4. In addition the issue of group sequential procedures for interim analyses was discussed in Chapter 10. In this chapter, we will focus on some remaining issues.

In Section 12.2, we will discuss baseline comparability between treatment groups with respect to patient characteristics. The concept of the intention-to-treat dataset is elucidated

and compared with the evaluable dataset in Section 12.3. Statistical methods for adjustment of treatment effects by covariates are discussed in Section 12.4. Some statistical issues regarding multicenter studies, multiple comparisons such as subgroup analyses, and multiple endpoints are addressed in Sections 12.5 and 12.6. Various aspects with respect to data monitoring are provided in Section 12.7. Section 12.8 illustrates the concept for utilization of genomic information for evaluation of the efficacy of a medicine under investigation. In Section 12.9 some final remarks are given.

12.2 BASELINE COMPARISON

Baseline measurements are those collected during the baseline periods as defined in the study protocol. Baseline usually refers to as measurements obtained at randomization and prior to treatment. Sometimes, measurements obtained at screening are used as baselines. Also, baselines are not always restricted to pretreatment measurements. For example, for complicated designs such as the enrichment design for the assessment of Tacrine in the treatment of Alzheimer's disease and CAST for arrhythmia, there may have several phases with the primary clinical endpoints evaluated at two successive phases. Then the measurements that would be used as the baseline for the entire trial or for a particular phase of the study should be precisely stated in the study protocol and in the final clinical report.

Basically the objectives for the analysis of baseline data are threefold. First, the analysis of baseline data provides a description of patient characteristics of the target population to which statistical inference is made, along with useful information on whether the patients enrolled in the study are representative of a targeted population according to the inclusion and exclusion criteria of the trial. Second, since baseline data measure the initial patient disease status, they can serve as reference values for the assessment of the primary efficacy and safety clinical endpoints evaluated after the administration of the active treatment. Finally, the comparability between treatment groups can be assessed based on baseline data to determine potential covariates for statistical evaluations of treatment effects.

In clinical trials the baseline data usually consist of demographic data such as age, gender, or race; initial disease status as evaluated by the primary efficacy, safety endpoints, and other relevant data; and medical history. The ICH *Guideline on Structure and Contents of Clinical Study Reports* requires that baseline data on demographic variables and some disease factors be collected and presented. These disease factors include (1) specific entry criteria, duration, stage and severity of disease and other clinical classifications and subgroupings in common usage or of known prognostic significance, (2) baseline values for critical measurements carried out during the study or identified as important indicators of prognosis or response to therapy, (3) concomitant illness at trial initiation, such as renal disease, diabetes, and heart failure, (4) relevant previous illness, (5) relevant previous treatment for illness treated in the study, (6) concomitant treatment maintained, (7) other factors that might affect response to therapy (i.e., weight, renin status, and antibody level), (8) other possibly relevant variables (e.g., smoking, alcohol intake, and special diets), and for women, menstrual status and data of last menstrual period if pertinent for the study.

Note that in most clinical trials, it is tempting to collect more baseline information than needed. It is suggested that we only focus on the baseline or prognostic variables that might affect the response to be treatment.

The effectiveness and safety of the treatment is usually assessed by the change from the baseline of some primary clinical endpoints. They change can be either the absolute

change from the baseline, which is defined as the difference between the post-treatment value and the baseline value, or the relative change such as the percent change from the baseline, which is the absolute change from the baseline divided by the baseline value multiplied by 100. The change from the baseline, which can be negative or positive, is symmetric about zero. The percent change from the baseline is in fact the ratio of the post-treatment value to the baseline value minus 1. Thus, the range of the percent change from the baseline is from -100 to infinity. As a result, the percent change from the baseline is not symmetric about 0 nor 1. If the post-treatment and the baseline values are positively correlated, then the variability of the change from the baseline will be smaller than that of the raw value. Therefore, the change from the baseline measures the alteration caused by the treatment. In addition, it may provide a more precise statistical inference of the treatment effects because of a possible reduction of variability. If the distribution of a clinical endpoint is approximately normal, then the distribution of the change from the baseline for this endpoint is also normal. However, the distribution of the percent change from the baseline is not normal; rather it is a Cauchy distribution whose moments such as mean and variance do not exist. As a result, no inference about the mean or variance can be made. In this case it is recommended that nonparametric methods be employed to obtain inference of the treatment effect based on the median.

As was indicated above, the baseline data are often used as reference values for assessing changes in disease status after administration of the treatment. In practice, multiple baseline measurements are often obtained not only to assess variability but also to evaluate the stability of the disease status before the administration of active treatment. However, the reasons for multiple baseline measurements are manifold (Carey et al., 1984; Frick et al., 1987; Manninen et al., 1988). For example, some investigators feel that due to various causes, a single baseline measurement might not provide reliable assessments for some critical disease characteristics. On the other hand, in antiarrhythmia trials, multiple baseline measurements are evaluated during the washout period to ensure that patients are free of previous antiarrhythmia medications. As a result, some of multiple baseline measurements evaluated at the end of the washout period can provide an actual description of the initial disease status for each patient. In addition the FDA guideline for the studies of benign prostatic hyperplasia (Boyarsky and Paulson, 1977) suggests that baseline measurements be collected in a placebo run-in period of at least 28 days. The purpose in collecting multiple baseline measurements during the placebo run-in period is to assess the placebo effect by some important clinical endpoints and to evaluate the stability of the disease status.

However, when there are more than one baseline measurement for a particular endpoint, the question arises *whether one of the measurements or the average of all or part of the measurements should be used as a single baseline value*. For example, in antihypertensive trials it is customarily to use the average of two or three measurements of systolic and diastolic blood pressures taken within the prespecified window of the baseline period as the baseline value without formal statistical justification. The question regarding which measurements should be used as baseline is subjective to clinical judgment, which should be addressed either in the study protocol or prior to the initiation of the trial based on the information from previous or related trials and pilot studies. In some cases, however, it may be of interest to establish the stability of the disease state or to evaluate the placebo effect through the placebo run-in period. In this case multivariate statistical procedures are useful. In practice, for multiple baseline measurements, it is suggested that valid statistical methods be used to justify a single baseline, such as the average, by combining baseline measurements.

Let Y_{hij} be the response of a clinical endpoint evaluated at the j th time point during the baseline period for the i th subject receiving the h th treatment; $h = 1, \dots, H$, $i = 1, \dots, n_i$, $j = 1, \dots, p$, and $n_i > p + 1$. Since baseline measurements are obtained prior to treatments, it is not anticipated that there are differences among treatment groups (due to the treatment). The following model is commonly used for investigation of the time effect of baseline measurements by suppressing the subscript h for treatment:

$$Y_{ij} = \alpha_j + e_{ij}, \quad j = 1, \dots, p, \quad i = 1, \dots, N, \quad (12.2.1)$$

where α_j is the average of clinical endpoint at time point j , e_{ij} is the random error associated with Y_{ij} , and $N = n_1 + \dots + n_H$.

Let the $p \times 1$ vector $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{ip})$ be the p measurements of the clinical endpoint measured at p time points during the baseline period. Also, let $\mathbf{e}_i = (e_{i1}, \dots, e_{ip})$ be the corresponding error vector. Assume that \mathbf{e}_i follow a multivariate normal distribution with mean vector 0 and a nonsingular covariance matrix Σ . Then the Hotelling T^2 can be used to investigate the trend and placebo effect for multiple baselines (Morrison, 1976). Let \mathbf{C} be a $(p-1) \times p$ matrix with the j th row vector \mathbf{c}'_j of orthogonal contrast such that

$$\mathbf{1}'_p \mathbf{c}_j = 0 \quad \text{and} \quad \mathbf{c}_j' \mathbf{c}_{j'} = 0 \quad \text{for } j \neq j',$$

where $\mathbf{1}_p$ is $p \times 1$ column of 1. Then the hypothesis of no overall trend effect due to time is rejected at the α th level of significance if the test statistic

$$\begin{aligned} T^2 &= \frac{N-p+1}{(n-1)(p-1)} \mathbf{N} \bar{\mathbf{Y}}' \mathbf{C}' (\mathbf{C} \mathbf{S} \mathbf{C}')^{-1} \mathbf{C} \bar{\mathbf{Y}} \\ &> F(\alpha, p-1, N-p+1) \end{aligned} \quad (12.2.2)$$

where $\bar{\mathbf{Y}}$ is the sample means of $\mathbf{Y}_1, \dots, \mathbf{Y}_N$, \mathbf{S} is the sample covariance matrix from $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ with $N-1$ degrees of freedom, and $F(\alpha, p-1, N-p+1)$ is α th upper quantile of an F distribution with $p-1$ and $N-p+1$ degrees of freedom.

Since the Helmert transformation compares a measurement of the baseline value at a particular time point to the average of baseline values at subsequent time points, it can be used to evaluate the placebo effect and determine the time point at which the baseline stabilizes (Searle, 1971). For the Helmert matrix, the j th row of \mathbf{C} is constructed such that the first $j-1$ elements are 0, the j th element is 1, and rest of the elements are $1/(p-j)$, $j = 1, \dots, p-1$. If the time point where the baselines stabilize can determined by the Helmert transformation, it is not only possible to estimate the placebo effect but also to combine the stabilized multiple baseline measurements into a single baseline. Note that if there is no time effect for the stabilized baseline measurements, model (12.2.1) can be rewritten as

$$Y_{ij} = \alpha + e_{ij}, \quad j = 1, \dots, p, \quad i = 1, \dots, N. \quad (12.2.3)$$

Under model (12.2.3), \mathbf{Y}_i follows a multivariate normal distribution with mean vector $\alpha \mathbf{1}_p$ and covariance matrix Σ . The criterion used to summarize the stabilized multiple baseline measurements of a clinical endpoint is to find the best linear unbiased estimate of α . In addition, since Y_{i1}, \dots, Y_{ip} represent p unbiased estimators of a scalar parameter α and covariance matrix Σ , the best linear unbiased estimator for α based on Y_{i1}, \dots, Y_{ip} by the

generalized least squares procedure (GLS), under model (12.2.3), is given by (O'Brien, 1984).

$$X_i = \frac{\mathbf{1}'_p \sum^{-1} \mathbf{Y}_i}{\mathbf{1}'_p \sum^{-1} \mathbf{1}_p} \quad (12.2.4)$$

with variance

$$V_X = \left(\mathbf{1}'_p \sum^{-1} \mathbf{1}_p \right)^{-1}.$$

Some investigators (e.g., Carey et al., 1984; Frick et al., 1987) suggest the use of a simple average to combine with stabilized multiple baseline measurements, which is given by

$$Z_i = \frac{\mathbf{1}'_p \mathbf{Y}_i}{\mathbf{1}'_p \mathbf{1}_p}, \quad (12.2.5)$$

with variance

$$V_Z = \frac{\mathbf{1}'_p \sum^{-1} \mathbf{1}_p}{p^2}.$$

Although the covariance matrix Σ is usually unknown, it can be estimated by its consistent sample estimator for large samples. Let $\mathbf{S}_1, \dots, \mathbf{S}_H$ be the within-group sample covariance matrices from H treatment groups. Then an unbiased consistent estimator of Σ can be obtained as

$$\mathbf{S}_p = \frac{1}{N - H} [(n_1 - 1)\mathbf{S}_1 + \dots + (n_H - 1)\mathbf{S}_H].$$

As a result the estimated generalized least square (EGLS) estimator of α is given by

$$x_i = \frac{\mathbf{1}'_p \mathbf{S}_p^{-1} \mathbf{Y}_i}{\mathbf{1}'_p \mathbf{S}_p^{-1} \mathbf{1}_p}, \quad (12.2.6)$$

where x_i is asymptotically unbiased and follows an asymptotic distribution such as that of (12.2.4). Unbiased consistent estimators of variances of x_i and Z_i are given by, respectively,

$$v_x = \frac{N - H}{N - p - 1} (\mathbf{1}'_p \mathbf{S}_p^{-1} \mathbf{1}_p)^{-1}$$

and

$$v_Z = \frac{\mathbf{1}'_p \mathbf{S}_p^{-1} \mathbf{1}_p}{p^2}.$$

Example 12.2.1 A parallel two-group randomized phase II study with 22 patients per treatment group was conducted to investigate the efficacy and safety of a newly developed pharmaceutical entity, compared to a placebo, in suppressing ventricular arrhythmia. One important safety endpoint in this trial was the supine heart rate (beats/minutes). Baseline measurements were obtained daily prior to the morning (placebo) dose during a four-day placebo run-in period. The sample means and pooled within-group sample covariance matrix are given in Table 12.2.1.

Table 12.2.1 Daily Mean Heart Rate (Beats per Minute) and the Pooled Within-Group Covariance Matrix

	Day 1	Day 2	Day 3	Day 4
Placebo				
N = 22	74.2	73.9	74.1	73.6
Treatment				
N = 22	75.0	73.5	73.1	73.8
Covariance				
Matrix ($df = 42$)				
Day 1	140.2	70.2	66.3	88.2
Day 2		110.8	45.6	47.5
Day 3			73.3	59.3
Day 4				92.9

To detect the placebo effect and determine the time point where the mean heart rate stabilizes, a preliminary analysis can be performed by the following 3×4 Helmert matrix:

$$\mathbf{C} = \begin{pmatrix} 1 & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \\ 0 & 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

The corresponding T^2 is 0.23 with a p -value of 0.87 (3 and 41 d.f.). Since we failed to reject the null hypothesis at the α level of significance, we can conclude that there is no placebo effect, and we do not pursue individual hypotheses about the time point where the heart rate stabilizes. Similar results for time trends can be tested based on the following linear contrasts:

$$\mathbf{C} = \begin{pmatrix} -3 & -1 & 1 & 3 \\ 1 & -1 & -1 & 3 \\ -1 & 3 & -3 & 1 \end{pmatrix}.$$

Since there is no placebo effect during the placebo run-in period, an attempt is made to construct a combined baseline heart rate for each patient from the four baseline measurements. The estimated generalized least square estimate of the combined baseline heart rate is given by

$$x_i = -0.1764Y_{i1} + 0.3027Y_{i2} + 0.5478Y_{i3} + 0.3259Y_{i4}.$$

The simple average of the four baseline heart rates, Z_i , uses equal weights for the combined baseline, while the EGLS procedures uses weights proportional to the row totals of the inverse of the pooled within-group sample covariance matrix. The estimated variances of x_i and Z_i are 66.32 and 73.21. Therefore, with respect to this dataset, the estimated variance of the combined baseline heart rate by the EGLS procedure is reduced about 9% compared to that of the simple average.

12.3 INTENTION-TO-TREAT PRINCIPLE AND EFFICACY ANALYSIS

In clinical trial development, despite a thoughtful and well-written study protocol, deviations from the protocol may be encountered over the course of a trial. For example, some patients randomly assigned to receive one of the treatments may be found not meeting the inclusion and/or exclusion criteria after the completion of the trial due to the reason that some laboratory evaluations (on which inclusion criteria are based) are not available at the time of randomization and drug dispensation. In some situations patients might receive wrong treatment due to a mix-up in randomization schedule. Some patients who received the assigned treatment might be switched to receive the other treatment in an emergency or after deterioration of the patient's disease status. In addition, for every clinical trial, it is likely that patients will withdraw from the study prematurely before the completion of the trial due to various reasons. For patients who complete the study, they might miss some scheduled visits. Another example is patient compliance to the treatment regimen. As a result a legitimate question is which patients should be included in the analysis for a valid and unbiased assessment of the efficacy and safety of the treatment.

Before we address the question, it is helpful to review the concept of randomization discussed in Chapter 4. Recall that randomization does not guarantee equal distributions of baseline variables such as demographic and patient characteristics among treatment groups. It, however, does provide an unbiased comparison between treatment groups. Cornell (1990) pointed out that randomization can avoid bias not only with respect to clinical endpoints, which can be readily identified, measured, and controlled in advance, but also with respect to those endpoints not measured and perhaps not yet known to be influential. In addition, as was indicated in Chapter 4, the random assignment of patients to treatment groups is the key to a valid statistical inference. However, as pointed out by Lachin (2000), randomization alone is not sufficient to provide an unbiased statistical inference. The sufficient conditions for a valid unbiased statistical inference include:

1. Missing data of randomized patients do not jeopardize unbiased comparisons between treatment groups.
2. Trial outcomes are assessed in a uniform and unbiased manner for all patients in the study.

In practice, the first condition can be achieved by the *intention-to-treat principle* (ITT principle). The ITT principle is described in the ICH E9 guidance entitled, *Statistical Principles for Clinical Trials* (1998). In the ICE E9 guidance, the ITT principle is referred to as "the principle that asserts that the effect of a treatment policy can be best assessed by evaluating on the basis of the intention to treat a subject (i.e., the planned treatment regimen) rather than the actual treatment given." It follows that all subjects allocated to a treatment should be followed up, evaluated, and analyzed as members of that group regardless of their compliance with the planned treatment. On the other hand, the second condition can be achieved through the blinding technique introduced in Chapter 4. Note that it is very important that blinding be enforced to all personnel involved during the study. The personnel include but are not limited to patients, clinicians, outcome assessors research nurses, trial pharmacists, and laboratory analysts.

Sackett et al. (1991) and Spilker (1991) report on a clinical trial that compares surgical therapy with medical therapy in treatment of patients with bilateral carotid stenosis. This

study enrolled 167 patients. Among these patients, 94 were randomized to the surgical group and 73 were randomly assigned the medical therapy. Sixteen patients who either had stroke or died during initial hospitalization were not included in the analysis. Fifteen of the 16 excluded patients were randomized to the surgical group. Consequently, a statistically significant number of patients in surgical therapy was excluded in the evaluable analysis (p -value = 0.0011 based on a two-sided Fisher's exact test). The rate of occurrence of subsequent transient ischemic attack, stroke, or death for the evaluable analysis is 54.4% in the surgical therapy as compared to that of 73.6% in the medical therapy. This indicates that there is a statistically significant reduction of risk 27% in the surgical therapy (p -value = 0.018 based on two one-sided Fisher's exact test). However, the intention-to-treat analysis, which includes all randomized patients, the reduction of risk is estimated as low as 16%, which is not statistically significant at the 5% level of significance (p -value = 0.10 based on the Fisher's two-sided test). These results are summarized in Table 12.3.1. The statistically significant reduction of risk in subsequent transient ischemic attack, stroke, or death of the surgical therapy observed from the evaluable analysis is a typical example of a biased inference and may be misleading because of the exclusion of patients from analysis.

Table 12.3.1 discussed above illustrates an excellent example of a biased inference from a reduced analysis set by treatment-related exclusion of patients. Therefore, from the definition, the ITT principle implies that the analysis should include all randomized subjects and their data should be analyzed according to their planned randomized treatment. In other words, the ITT principle requires complete follow-up of all randomized subjects for study outcomes. The set of subjects included in the analysis that is as close as possible to the ITT principle is referred to as the *full analysis set* (ICH E9, 1998). In practice, however, the full compliance of the ITT principle is difficult to accomplish. For example, after randomization, subjects may be found to violate some major inclusion/exclusion criteria, patients may fail to take any assigned treatment drug, or patients may just disappear and hence do not provide any postrandomization data. As a result, the full analysis set may not include these subjects. Consequently, potential bias may be introduced from the exclusion

Table 12.3.1 Summary of Surgical Versus Medical Therapy in Treatment of Bilateral Carotid Stenosis

	Therapy		
	Surgical	Medical	p -Value
<i>I. Patients excluded</i>			
N	94	73	0.0011
Number excluded	15 (16.0%)	1 (1.4%)	
<i>II. Evaluable analysis</i>			
N	79	72	0.018
TIA, stroke or death	43 (54.4%)	53 (73.6%)	
<i>III. Intention-to-treat analysis</i>			
N	94	73	0.10
TIA, stroke or death	58 (61.7%)	54 (74.8%)	

Note: TIA: Transient ischemic attack; p -values are obtained from the two-tailed Fisher's exact test.

Source: Adapted from Sackett et al. (1991) and Spilker (1991).

of these subjects. Therefore, it is suggested that a discussion be provided to address this issue as fully as possible. For example, whether the entry criteria were measured before randomization and detection of violation of entrance criteria were made objectively and uniformly for all subjects should be fully addressed. In addition, whether failure to take study medication or failure to return after the randomization visit has anything to do with or is influenced by knowledge of treatment assignment should also be addressed.

In general, the ITT analysis is a conservative approach with better capability of providing an unbiased inference about the treatment effect. The ITT analysis can reflect real clinical practice better than any other analyses. In order to provide an unbiased and valid inference based on the ITT analysis, the study protocol should include dropout rates, rates of poor compliance, patterns of missing values, and other related variables as response variables. These response variables should be compared between the treatment groups. In addition interactions between treatment groups and noncompleters (or completers) should also be examined with respect to demographic variables and baseline disease characteristics. The CPMP Working Party on Efficacy of Medicinal Products suggests that at least demographic and baseline data on the disease status in the wider population screened for entry, or enrolled into a screening phase of the study prior to randomization, should be summarized to provide information about the numbers and characteristics of excluded patients, both eligible and ineligible, together with their reasons for exclusion, in order to guide assessment of the potential practical impact on the study results (European Commission, 1990, 1994).

For the ITT analysis, serious bias can be introduced if no data are collected for the patients who discontinued the study prematurely. Therefore it is recommended that the primary clinical endpoints of the trial be evaluated at the time of withdrawal. In many clinical trials the method of the last observation carried forward (LOCF) is applied to patients who withdrew prematurely from the study with no additional follow-up data after discontinuation. This approach, however, is biased if the withdrawal of patients is treatment related. In principle, the ITT analysis is performed according to the random allocation of treatment regardless of the real treatment that patients actually receive. If the proportion of mix-up in treatment assignments exceeds a specified level, the ITT analysis, nor any other analysis, will no longer be valid for the interpretation of the results. This is the fatal consequence of poor conduct of randomization codes, packaging, and dispensation of the study drugs. If this situation does occur, the trial has little value in providing clinical evidence for the efficacy and safety of the study drug.

Thus far, we only focus on the analysis aspect of the ITT principle. As mentioned before, however, a biased inference may occur in the presence of missing data, which may result from premature discontinuation of assigned treatment or withdrawal of subjects before their scheduled completion of the study or poor adherence of subjects to the study schedule. Note that statistical analysis based on the full analysis set does not fully comply with the ITT principle due to the exclusion of the subjects with violation of entry criteria or with no post-randomization data. As a result, this analysis is not a true intention-to-treat analysis, but it is another subset analysis. The ITT principle, therefore, requires *complete* follow-up of all randomized subjects for study outcomes, and hence, an unbiased comparison among treatments can be made based on complete data from all randomized subjects. This concept in the design of clinical trials is referred to as the ITT design. Lachin (2000) defines the ITT design, wherein all subjects are followed until death or the end of trial, until the outcome is reached regardless of whether the subject is still receiving or complying with the assigned treatment. It follows that premature discontinuation of assigned treatment is different from premature withdrawal of subjects from the study. Subjects may discontinue their randomly

assigned treatment due to various reasons such as safety. However, as Lachin (2000) pointed out, premature discontinuation of treatment should not lead to premature withdrawal of subjects from the study. In addition, regardless of premature discontinuation of assigned treatments, all scheduled assessment and evaluation should be performed on all subjects until the death of the subject or the scheduled conclusion of the study. Only in this manner can a true ITT analysis be performed.

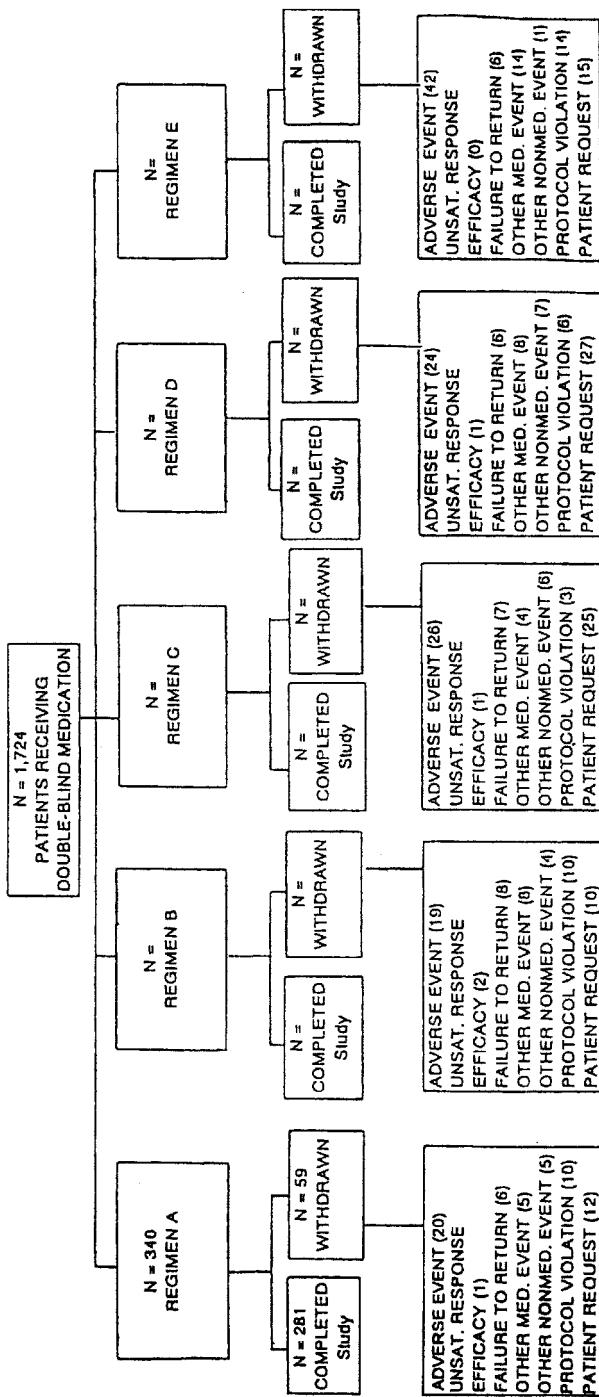
In addition to the ITT analysis, many subsets of the ITT dataset may be constructed for efficacy analysis. These subsets may include (1) all patients with any efficacy observations or with a certain minimum number of observations, (2) only patients completing the study, (3) all patients with an observation during a particular time window, and (4) only patients with a specified degree of compliance. Analyses based on these subsets are referred to as *evaluable analyses*. The patient population for the evaluable analysis is called the *per-protocol* or *efficacy* patient population. The criteria for inclusion of patients in the evaluable analysis should be specified in the study protocol in advance. It should not be established after the data are collected and randomization codes are unblinded. The criteria for the evaluable analysis usually include the following characteristics:

1. Satisfaction of a prespecified minimal length of exposure to the treatment of the investigated therapy.
2. Provision of data on primary clinical endpoints at prespecific and relevant scheduled time points.
3. Satisfactory compliance with treatment.
4. No major protocol violations or deviations including the inclusion or exclusion criteria, incorrect randomization, concomitant medications.

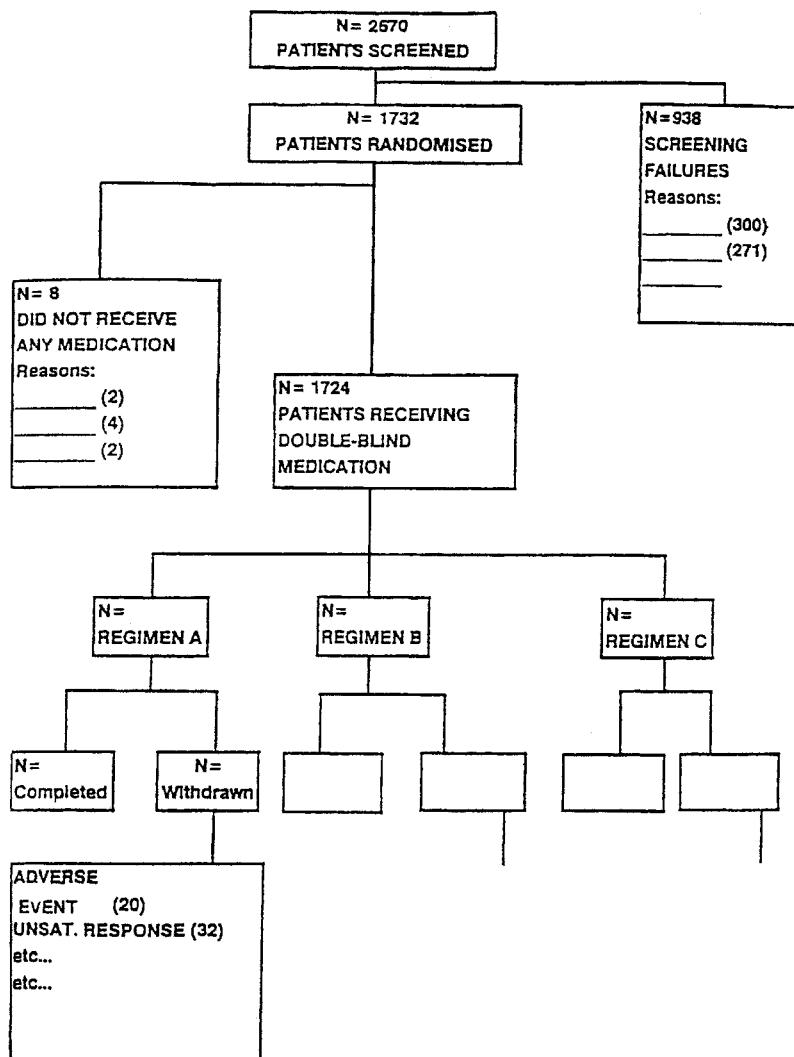
Note that the above criteria should be specified in the study protocol. The definition of protocol violation should also be specified in the study protocol prior to the initiation of the trial. In addition the nature and reasons for protocol violation should be reviewed, described, decided, and documented by blinded data coordinators, medical monitor and other personnel before the randomization codes are broken and the database is locked for analysis. The evaluable analysis is usually considered the preferred analysis by the sponsor because, in general, it maximizes the probability of showing the efficacy of the drug under study. The evaluable analysis, however, is much more vulnerable to bias than the ITT analysis because of inherent subjectivity for exclusion of patients from the evaluable analysis as demonstrated by the example given in Table 12.2.1. For clinical trials with comparative controls or placebos, the ITT analysis will give a less optimistic estimate for the efficacy, since the inclusion of noncompleters will generally dilute the treatment effect. Lachin (2000) pointed out that the per-protocol analysis not only inflates the type I error rate up to 50% or higher even when the null hypothesis is true, but also it provides less power than the ITT analysis. On the other hand, if the objective of the trial is to show therapeutic equivalence between treatment groups, the ITT analysis is still less biased than the evaluable analysis, but it is no longer conservative in the context of equivalence trials. The results obtained from the ITT analysis should be interpreted with extreme caution.

From the above discussion, it is helpful to provide a disposition of all patients who entered the study so that the patients in the datasets for the ITT and evaluable analyses can be clearly defined. The principle is to account for every single patient participated in the study. Both the FDA and ICH guidelines on clinical reports require that the numbers of patients who are randomized, and complete each phase of the study be provided, as well as

Table 12.3.2 Example of Tables for Disposition of Patients



Source: *Guideline on Structure and Contents of Clinical Study Reports* (ICH, 1996).

Table 12.3.3 Example of Tables for Disposition of Patients

Source: Guideline on Structure and Contents of Clinical Study Reports (ICH, 1996).

the reasons for all post-randomization discontinuation. Tables 12.3.2 and 12.3.3 give two examples of the summary data on patient disposition as provided in the ICH *Guideline on the Structure and Content of Clinical Study Reports* (ICH, 1996). The FDA guideline also suggests, as illustrated in Table 12.3.4, that there be provided summaries of the number of the patients who entered and completed each phase of the study, or each week/month of the study. In addition a listing should be provided as shown in Table 12.3.5, by center and treatment group, for the patients who discontinued from the trial after enrollment, with the information on patient identifier, demographic characteristics, the specific reasons for discontinuation, the treatment and dose level, and the duration of the treatment before discontinuation. After the status of each patient during the trial is accounted for, we need to decide

Table 12.3.4 Disposition of Patients

	STUDY # (Data Set Identification)				
	Number of Patients Completing Each Period of Study				
	Randomized	Treated	Week 1	Week 2	Week 4
Test drug	#	#(%)			
Active control					
Placebo	_____	_____	_____	_____	_____
Total	_____	_____	_____	_____	_____
Comparability test (<i>p</i> -value)	_____	_____	_____	_____	_____

Source: U.S. FDA Guideline for the Format and Content of the Clinical and Statistical Sections of an Application (1988).

on the dataset for efficacy, in particular, which patients should be included in the efficacy analysis. These information should be provided in the clinical report. Table 12.3.6 gives a listing of the patients and visits excluded from an efficacy analysis. The FDA guidelines also require the investigator to provide a summary table of the number of patients excluded from the efficacy analysis by reason and week or phase (see Table 12.3.7). The primary efficacy analysis should be based on all randomized patients according to their random assignments of treatment. If the preferred efficacy analysis by the sponsor is based on a reduced subset of the intention-to-treat dataset, any discrepancies and inconsistencies between the two analyses should be explained and clarified in the report.

12.4 ADJUSTMENT FOR COVARIATES

For assessment of the efficacy and safety of a drug product, it is not uncommon that primary clinical endpoints are affected by some factors such as the demographic variables of age, gender, and race and/or patient characteristics such as disease severity, concomitant medications, and medical history. These factors are referred to as covariates and are also known as confounding factors, prognostic factors, or risk factors. In clinical trials, we might identify some covariates of an impact on the clinical outcomes so that these covariates can be measured during the trial. In practice, it is common not to collect information on covariates that may be influential and yet unknown at the planning stage of the trial. If patients are randomly assigned to receive treatments, as indicated in Chapter 4, the estimated treatment effect is asymptotically free of the accidental bias induced by omission of one or more covariates. In other words, a randomized trial is asymptotically free of covariate imbalance, even for unknown and unmeasured covariates. Covariate balance can typically be reached when the sample size tends to infinity. In practice, however, the sample size is finite; consequently the estimated treatment effect is biased if there is an imbalance in one or more covariates.

As an example, let us consider a antihypertensive trial that compares a test drug against a placebo. Let Y_{ij} and X_{ij} be the diastolic blood pressure (mmHg) measured at the end of the study and its corresponding baseline value for j th patient and the i th treatment; $j = 1, \dots, n_i$, $i = T, P$. The analysis of covariance model is given in (8.5.1) is

$$Y_{ij} = \mu + \tau_i + \beta(X_{ij} - \bar{X}_{..}) + \varepsilon_{ij}, \quad (12.4.1)$$

Table 12.3.5 Listing of Patients Who Discontinued Therapy

Center:	Study #						Reason for Discontinuation	
	Data Set Identification			Treatment				
Treatment	Patient #	Sex	Age	Last Visit	Duration	Dose	Concomitant Medication	Reason for Discontinuation
Test drug/Investigational product								
Treatment	Patient #	Sex	Age	Last Visit	Duration	Dose	Concomitant Medication	Reason for Discontinuation
Active control/Compactor								
Treatment	Patient #	Sex	Age	Last Visit	Duration	Dose	Concomitant Medication	Reason for Discontinuation
Placebo								

(Repeat for other centers)

Source: The U.S. FDA Guideline for the Format and Contents of the Clinical and Statistical Sections of an Application (1988).

*The specific reaction leading to discontinuation.

Table 12.3.6 Listing of Patients and Observations Excluded from Efficacy Analysis

STUDY # (Data Set Identification)						
Center: Treatment	Patient #	Sex	Age	Observation	Excluded	Reason(s)
Test Drug/ Investigational Product						
Treatment	Patient #	Sex	Age	Observation	Excluded	Reason(s)
Active Control/ Comparator						
Treatment	Patient #	Sex	Age	Observation	Excluded	Reason(s)
Placebo						
(Repeat for other centers)						
Reference Tables						
Summary:						

Source: U.S. FDA *Guideline for the Format and Content of the Clinical and Statistical Sections of an Application* (1988).

where $\bar{X}_{..}$ is the overall average of the baseline values. The parameter of interest is the treatment effect, which is given by

$$\theta = \tau_T - \tau_P.$$

We denote \bar{Y}_i and \bar{X}_i as the average of the diastolic blood pressures and the corresponding baseline values of the i th treatment, $i = T, P$. If the baseline diastolic blood pressure is

Table 12.3.7 Number of Patients Excluded from Efficacy Analysis

STUDY # (Data Set Identification)			
Test Drug	N =		
	Week		
Reason	1 2 4 8		
_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____
_____	_____	_____	_____
Total	_____		

Note: Similar tables should be prepared for the other treatment groups.

Source: U.S. FDA *Guideline for the Format and Content of the Clinical and Statistical Sections of an Application* (1988).

ignored, then an intuitive estimate of the treatment effect is the difference in the unadjusted treatment average between the test drug and the placebo, namely

$$\hat{\theta}^* = \bar{Y}_T - \bar{Y}_P. \quad (12.4.2)$$

The expected value of $\hat{\theta}^*$ is

$$E(\hat{\theta}^*) = \tau_T - \tau_P + \beta(\bar{X}_T - \bar{X}_P). \quad (12.4.3)$$

From (12.4.3) we note that in addition to the true treatment effect, $E(\hat{\theta}^*)$ contains a covariate. Therefore, unless the baseline diastolic blood pressure is balanced between the test and the placebo groups, namely $\bar{X}_T = \bar{X}_P$, the estimated treatment effect (12.4.2) is biased. We define

$$\bar{Y}_i^* = \bar{Y}_{i\cdot} - \hat{\beta}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}) \quad (12.4.4)$$

as the adjusted treatment average for the i th treatment, $i = T, P$, where $\hat{\beta}$ is the least squares estimate (LSE) of the regression coefficient of the diastolic blood pressure at the end of the study on the corresponding baseline value. Since the LSE of β is unbiased, the expected value of the adjusted treatment average is given by

$$\begin{aligned} E(\bar{Y}_i^*) &= E[\bar{Y}_{i\cdot} - \hat{\beta}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})] \\ &= \mu + \tau_i + \beta(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}) - \beta(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}) \\ &= \mu + \tau_i. \end{aligned}$$

Let $\hat{\theta}$ denote the difference in adjusted treatment average between the test drug and the placebo. The expected value of $\hat{\theta}$ is given by

$$\begin{aligned} E(\hat{\theta}) &= E(\bar{Y}_T^* - \bar{Y}_P^*) \\ &= \tau_T - \tau_P. \end{aligned}$$

Since the expected value of $\hat{\theta}$ consists of only the treatment effect which is independent of the covariate, it is an unbiased estimator of the unknown but true treatment effect. As discussed above, if the covariates are not balanced, then the difference in the simple average between treatment groups will be biased for estimation of the treatment effect. Hence the covariates must be included in the statistical model for an unbiased estimate of the treatment effect.

In the case where covariates are not balanced between the treatment groups, to obtain a valid inference of the treatment effect, it is necessary to adjust for covariates that are statistically significantly correlated with the clinical endpoints. As was mentioned above, if a covariate is balanced, then the difference in the simple averages between treatments is an unbiased estimate for the treatment effect. This is equivalent to assuming the following one-way analysis of variance model:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}^*. \quad (12.4.5)$$

A comparison between models (12.4.1) and (12.4.5) reveals that the analysis of covariance is a combination of analysis of variance and regression analysis techniques. As a result, the error term ε_{ij}^* in model (12.4.5) not only includes the pure error term ε_{ij} but also the part of the regression of the response endpoint on the covariate. If a statistically significant correlation exists between the response endpoint and the covariate, then a significant portion of variability of the response endpoint can be explained by the covariate. As a result, the residual sum of squares obtained under the model (12.4.1) is much smaller than that under model (12.4.5) without the inclusion of the covariate. Although both models (12.4.1) and (12.4.5) yield unbiased estimators of the treatment effect, the precision of the estimated treatment effect under model (12.4.1) is better because the covariate helps to remove the variability of the response endpoint. In summary, an adjustment of covariates not only provides unbiased statistical inference but also increases precision of the statistical inference.

Note that the model (12.4.1) with covariates in statistical inference for estimation of the treatment effect assumes a common slope for both the test and placebo groups. The treatment effect at a particular value of the covariate is then the distance between the two lines at that value. Under the assumption of a common slope, the regression lines for the two groups in their relationship to the response endpoint and the covariate, which are parallel to each other, are shown in Figure 12.4.1. Since the distance between the two lines is the same for the entire range of the covariate, a common treatment effect can be estimated.

If there is an interaction between treatment and covariate, then the two regression lines will not have the same slope. In this case model (12.4.1) can be modified as

$$Y_{ij} = \mu + \tau_i + \beta_i(X_{ij} - \bar{X}_.) + \varepsilon_{ij}. \quad (12.4.6)$$

It follows that the two regression lines are not parallel any more. This implies that the treatment effect is different at different values of the covariate. Although an unbiased estimate can still be obtained at a particular value of the covariate, it is not possible to unbiasedly

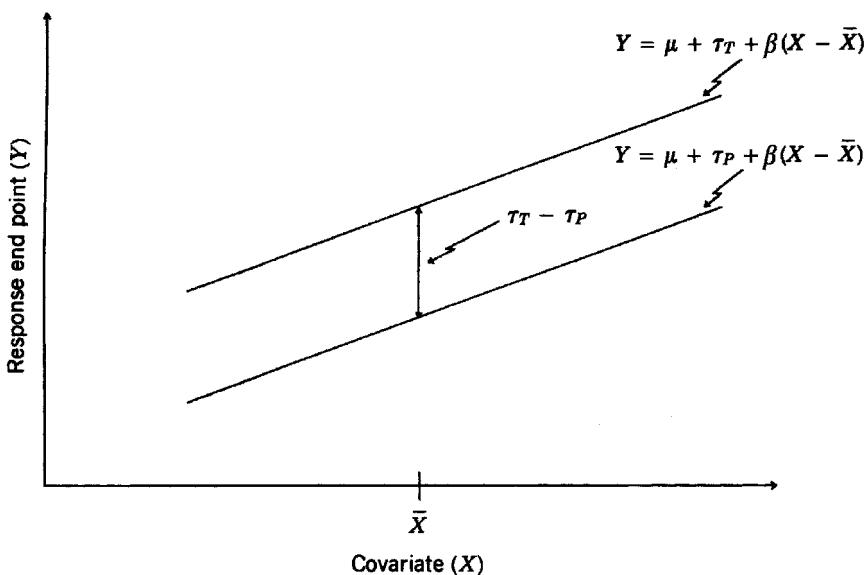


Figure 12.4.1 Adjustment for covariate in estimation of treatment effect. Case I: Common slope.

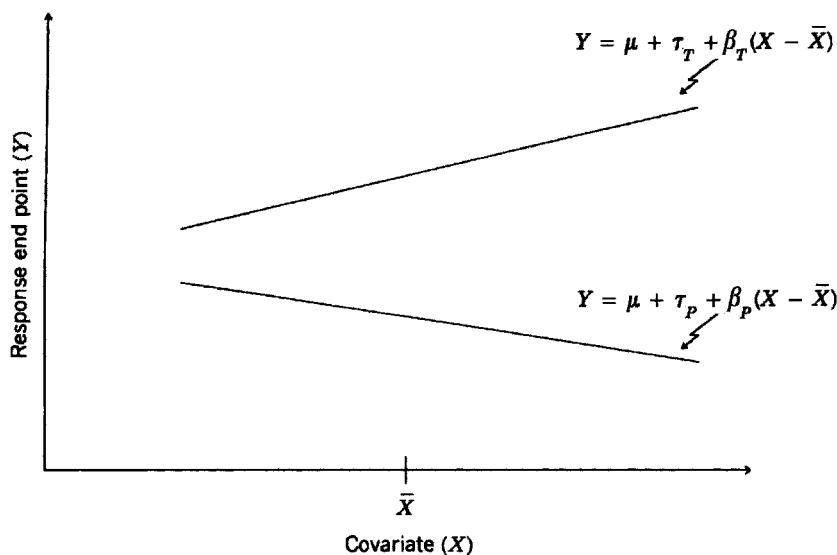


Figure 12.4.2 Adjustment for covariate in estimation of treatment effect. Case II: Different slopes, same direction but different magnitude.

estimate the common treatment effect over the entire range of the covariate. Figure 12.4.2 depicts a situation where the treatment effect has the same direction for the possible range of the covariate. The treatment effect increases as the value of the covariate increases. This might indicate that test drug is more efficacious for those patients with a large covariate value. On the other hand, the treatment effect might change signs at the different ranges of the covariate as shown in Figure 12.4.3. In other words, the test drug is worse than the placebo for patients with low covariate values. If interaction between the treatment and a covariate exists, a stratified analysis might be performed by dividing the entire range of covariates into several strata so that the treatment effect is homogeneous. This subgroup analysis will be discussed in this chapter.

Another key assumption in the adjustment of covariates of the treatment effect under model (12.4.1) is that the treatment should not affect the covariates. This assumption is easily satisfied by subject-specific covariates such as demographic variables and baseline disease characteristics that are measured only once prior to the initiation of the treatment and are time independent. Certain other covariates may be measured at every post-treatment visit when the primary clinical endpoints are evaluated. Typical examples of such covariates are heart rates and cholesterol levels in the antihypertensive trials or in a prevention trial of cardiovascular events, and CD4 in AIDS trial. Since these covariates are likely to be influenced by the treatment, interpretation of the estimated treatment effect after adjustment for the time-dependent covariates must be made with extreme caution. As a result the CPMP Working Party on Efficacy of Medicinal Products advises investigators not to adjust the primary analysis for covariates measured after randomization (European Commission, 1994). In addition, it is recommended that statistical methods that account for covariate imbalance between treatment groups be used to improve the precision of the inference.

Model (12.4.1) is a model for the adjustment of covariates with continuous endpoints. The blocked Wilcoxon rank sum test (Lehmann, 1975) can be used for categorical covariates when the normality assumption of continuous endpoints are questionable. The logistic

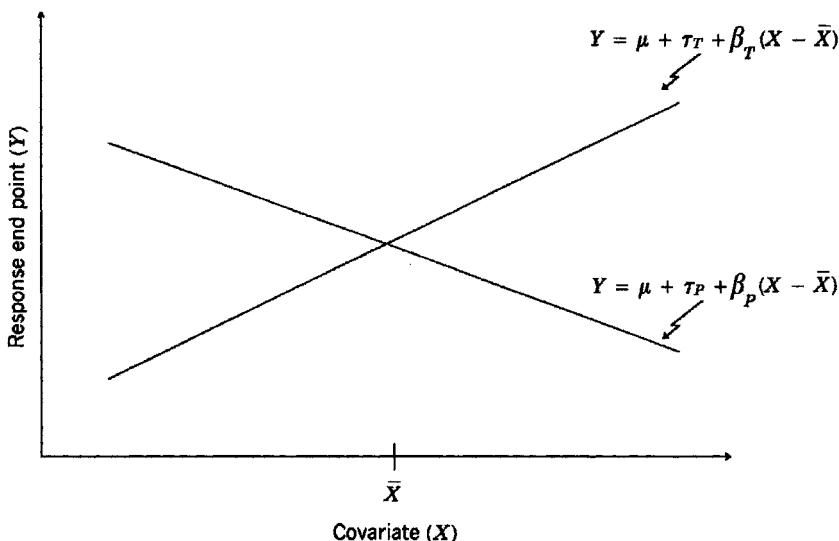


Figure 12.4.3 Adjustment for covariate in estimation of treatment effect. Case III: Different slopes, different direction with different magnitude.

regression or generalized estimation equations (GEE) discussed in Chapter 9 can be employed for the analysis of categorical endpoints. The adjustment for covariates with censored data can be handled through Cox's proportional hazard model, as described in Chapter 10. Since there are many possible covariates and statistical methods for adjustment, the FDA and ICH guidelines for clinical reports require that the selection of and adjustment of any covariate should be explained in the reports. In addition methods for adjustment, results of analyses, and supportive information should be fully documented along with other statistical methods. If the covariates or the methods for adjustment are different from those planned in the protocol, the results of the planned analyses should be presented and the discrepancies should be explained in detail. Examples of how the adjustment of covariates results should be presented are given in the FDA guideline, these examples are reproduced in Tables 12.4.1 and 12.4.2.

12.5 MULTICENTER TRIALS

A multicenter trial is a single study that is conducted simultaneously at more than one site (or center) according to a common protocol. A multicenter trial is not equivalent to separate *single-site* trials. Hence the data collected from different centers of a multicenter trial are intended to be analyzed as a whole. The history of multicenter trials can be traced back to early 1960s when Sir Bradford Hill designed his classical multicenter controlled trial of antihistamine, cortisone, and streptomycin (Hill, 1962).

There are many reasons for conducting multicenter trials. The most important rationale is probably that the framework of multicenter trials provides an efficient means of accruing sufficient numbers of patients in order to achieve a desired power within a predetermined time frame. If a trial is conducted at a single site, the resulting estimate of the treatment effect might be uniquely related to some characteristics, known or unknown, of the center.

Table 12.4.1 Summary of Results for Prognostic Factors for Analysis of Time to Treatment Failure

Prognostic Factor	Study 1 ^b	Study 2
Baseline performance status	<0.01	<0.01
Age	0.02	0.29
Disease-free interval	0.08	0.48
Adjuvant chemotherapy	0.53	0.01
Race	0.96	0.12
Estimated coefficient for treatment, beta ^a	-0.12	-0.27
Standard error of beta	0.12	0.13
p-value of treatment effect	0.30	0.03

e^{BETA} = estimated ratio of hazard rates
 (active:test)^a 0.89 0.76
 (95% confidence interval) (0.70, 1.12) (0.59, 0.99)

^aAdjusted for all prognostic factors whose p-value was <0.20 (significance at 0.20 level should not be taken as an FDA policy).

^bp-Value from final Cox regression model.

Source: U.S. FDA Guideline for the Format and Content of the Clinical and Statistical Sections of an Application (1988).

Thus the generalization of trial results can be very restricted. Multicenter trials provide a basis for a broad generalization of the trial results because the patients are recruited from a wider spectrum of the targeted population. In addition, the study drug is administrated under a broader clinical setting and practice, with a large number of investigators who will provide a more comprehensive clinical judgment concerning the value of the study drug.

Table 12.4.2 Effect of Prognostic Factors of the Comparison of Treatments Before and After Adjusting for Prognostic Factors

Efficacy Variable	Study 1	Study 2
<i>Time to treatment failure</i>		
<i>Estimated hazard ratio (T:A)</i>		
Unadjusted	0.90	0.81
(95% confidence limits)	(0.72, 1.14)	(0.63, 1.04)
p-value	0.39	0.09
<i>Estimated hazard ratio (T:A)</i>		
Adjusted	0.89	0.76
(95% confidence limits)	(0.70, 1.12)	(0.59, 0.99)
p-value	0.30	0.03
Factors adjusted	PS, AGE, INT ^a	PS, ADJ, RACE ^a

Note: T = test drug, A = active control.

^aPS = Baseline performance status, AGE = age, INT = disease-free interval, ADJ = adjuvant chemotherapy, RACE = race.

Source: U.S. FDA Guideline for the Format and Content of the Clinical and Statistical Sections of an Application (1988).

Table 12.5.1 Example for the Distribution of Patients by Center and Treatment—I

Center	Placebo	Low	Medium	High	Total
San Diego	16	19	17	18	70
Phoenix	15	17	18	14	64
Houston	15	18	20	16	69
Minneapolis	1	0	2	0	3
Total	47	54	57	48	206

Although a multicenter trial possesses the characteristics that (1) it is conducted under a single protocol, (2) it follows a set of pre-specified schedules and evaluation procedures, and (3) the data are processed under a centralized management system, patients recruited at different centers may be inherently subtly different due to the fact that (a) individual investigator's clinical practice and standard operating procedures might be consistent but not completely the same and (b) equipment might be differently calibrated. These issues and other known and unknown reasons may cause variation among centers. In addition, the number of patients recruited at each center is in fact a random number that cannot be controlled as desired. For example, the protocol might call for at least 20 patients at each center. It might turn out that distribution of the patient is quite heterogeneous across centers. Table 12.5.1 provides a tabulation of the number of patients for the example given in Section 2.3. This example presents the situation where a phase IIB multicenter trial with four treatment groups was conducted at a small number of centers. A large number of patients, namely more than 60, were randomized at three of the four centers. However, the other center only recruited three patients. As a result, at this center no patients were even randomized to the low- and high-dose groups. Table 12.5.2 presents another situation where a phase IIB multicenter trial with four treatment groups was conducted at a large number of centers. Although it was anticipated that 20 patients were recruited at each center, only 9 of the 25 centers reached the goal which constitutes 60% of 428 randomized patients. On the other hand, fewer than 15 patients were randomized at other 9 centers with a total of 97 patients (or 23% of 428 randomized patients). In addition two centers only randomized no more than three patients. For both examples, since at some centers some treatment groups did not have randomized patients, the result/interpretation of statistical analyses were not reliable.

For the analysis of a multicenter trial, Lewis (1995) posed the following questions which are helpful for the analysis of clinical data in a regulatory context:

1. Are some of the centers too small for reliable separate interpretations of the results?
2. Are some of the centers so big that they dominate the results?
3. Do the results at one or more centers look out of line with the others, even if not significantly so? Can the reason for this be established?
4. Do any of the centers show a trend in the *wrong* direction? Can this be explained? How can we restrict the use of a drug to the appropriate patients if we do not understand such trends?
5. If a treatment-by-center interaction is detected, is the trial validated?

Both the FDA and ICH guidelines require statistical tests for homogeneity across centers in order to detect possible quantitative or qualitative treatment-by-center interaction. For

Table 12.5.2 Example for the Distribution of Patients by Center and Treatment—II

Center	Placebo	100 mg	200 mg	300 mg	Total
1	5	5	5	6	21
2	3	4	3	4	14
3	6	6	6	6	24
4	3	2	2	3	10
5	5	5	5	5	20
6	4	3	4	3	14
7	9	9	9	9	36
8	7	6	7	6	26
9	9	9	9	9	36
10	6	6	5	6	23
11	2	2	3	3	10
12	1	0	1	1	3
13	5	5	4	5	19
14	9	9	8	8	34
15	5	5	4	5	19
16	2	2	2	2	8
17	4	5	4	5	18
19	4	5	5	5	19
20	4	3	3	3	13
21	9	9	9	9	36
22	2	4	4	3	13
24	2	3	3	2	10
25	0	1	1	0	2
Total	106	107	106	109	428

example, let Y_{ijk} be the reduction in diastolic blood pressure (mmHg) observed from patient k receiving treatment i at center j in a clinical trial that evaluates the efficacy and safety of a test drug with a placebo control at J centers in treatment of patients with mild to moderate hypertension; $k = 1, \dots, n_{ij}$, $j = 1, \dots, J$, $i = T, P$. It is assumed that Y_{ijk} approximately follows a normal distribution with population average μ_{ij} and variance σ^2 . Here we use a normal continuous endpoint. Similar concepts can be applied to non-normal continuous data, categorical data, or censored data from a multicenter trial. Descriptive statistics for site j are summarized in Table 12.5.3. Traditionally, data from multicenter trials can be analyzed by the two-way analysis of variance model with or without interaction, as discussed in Section 8.4. The model with interaction is given as

$$\begin{aligned} Y_{ijk} &= \mu_{ij} + \varepsilon_{ijk} \\ &= \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \end{aligned} \quad (12.5.1)$$

The analysis of variance model without interaction can be obtained from (12.5.1) by excluding the interaction term $(\tau\beta)_{ij}$

$$\begin{aligned} Y_{ijk} &= \mu_{ij} + \varepsilon_{ijk} \\ &= \mu + \tau_i + \beta_j + \varepsilon_{ijk}, \end{aligned} \quad (12.5.2)$$

where ε_{ijk} is assumed to follow a normal distribution with mean 0 and variance σ^2 .

Table 12.5.3 Descriptive Statistics for Site j of a Multicenter Trial

Statistics	Treatment		Difference
	Placebo	Test Drug	
N	n_{Pj}	n_{Tj}	
Mean	$\bar{Y}_{Pj.}$	$\bar{Y}_{Tj.}$	$d_j = \bar{Y}_{Pj.} - \bar{Y}_{Tj.}$
Standard deviation	s_{Pj}	s_{Tj}	s_j
Confidence Interval	CI_{Pj}	CI_{Tj}	CI_j
$s_j^2 = [(n_{Pj} - 1)s_{Pj}^2 + (n_{Tj} - 1)s_{Tj}^2]/(n_{Pj} + n_{Tj} - 2)$			
$CI_{ij} = \bar{Y}_{ij.} + (s_{ij}/\sqrt{n_{ij}})t(\alpha/2, n_{ij} - 1), i = T, P$			
$CI_j = (\bar{Y}_{Tj.} - \bar{Y}_{Pj.}) + (s_j/\sqrt{w_j})t(\alpha/2, n_{Pj} + n_{Tj} - 2); w_j = (1/n_{Pj}) + (1/n_{Tj})$			

At center j the treatment effect between the test drug and the placebo can be expressed as

$$\delta_j = \mu_{Tj} - \mu_{Pj}. \quad (12.5.3)$$

An unbiased estimator for δ_j can be obtained as

$$d_j = \bar{Y}_{Tj.} - \bar{Y}_{Pj.}, \quad (12.5.4)$$

with an estimate of its variance given by

$$\begin{aligned} v(d_j) &= s_j^2 \\ &= \frac{(n_{Pj} - 1)s_{Pj}^2 + (n_{Tj} - 1)s_{Tj}^2}{n_{Pj} + n_{Tj} - 2}, \quad j = 1, \dots, J. \end{aligned} \quad (12.5.5)$$

As a result, at site j , the null hypothesis of no treatment effect is rejected at the α level of significance if

$$|t_j| = \left| \frac{d_j}{s_j/\sqrt{w_j}} \right| > t(\alpha/2, n_{Pj} + n_{Tj} - 2), \quad j = 1, \dots, J, \quad (12.5.6)$$

or the $(1 - \alpha)100\%$ confidence interval of δ_j does not contain zero.

The objective for the analysis of clinical data from a multicenter trial is twofold. It is not only to investigate whether a consistent treatment effect can be observed across centers but also to provide an estimate of the overall treatment effect. These objectives, however, depend on (1) whether the center should be considered a fixed or a random factor, (2) what is the appropriate definition of the overall treatment effect, and (3) whether one should estimate the overall treatment in a manner similar to a stratified analysis. Table 12.5.2 illustrates the possibility that a multicenter trial might have a large number of centers, each with a relatively small number of patients. In addition, it is reasonable to consider the centers participating in a multicenter trial as a representative sample randomly selected from a population of centers. Chakravorti and Grizzle (1975) recommend that the center effect be considered

as random rather than fixed. Under the assumption that the center effect is random, model (12.5.1) becomes a mixed effect model. Under this model, Fleiss (1986b) indicates that the inference of the difference between the treatment averages can be made by a linear combination of the pooled variance s^2 and the interaction mean square $\text{MS}(AC)$, where

$$s^2 = \frac{\sum \sum (n_{ij} - 1)s_{ij}^2}{\sum \sum (n_{ij} - 1)}, \quad (12.5.7)$$

$$\text{MS}(AC) = \frac{1}{J-1} \sum \left(\frac{1}{w_j} \right) (d_j - \bar{d}^*)^2, \quad (12.5.8)$$

and

$$\bar{d}^* = \frac{\sum (1/w_j)d_j}{\sum (1/w_j)}.$$

Except for the balanced situation where the equal number of responses for the clinical endpoints observed for each treatment-by-center combination is proportional to its marginal (i.e., $n_{ij} = n_i n_j / n$ for all i, j), statistical methods can be very complicated in the analysis of mixed effects from a multicenter trial, and only approximate results are available (Searle, 1971; Fleiss, 1986b; Chkavorti and Grizzle, 1975; Mielke and McHugh, 1965). Hence, the analysis of the data from a multicenter trial under the assumption of a mixed effects model is unnecessarily perplexing.

In practice, a center is often selected based on the criteria that (1) it can recruit a minimum number of patients from the targeted population as specified by the sponsor, (2) the investigators at the center have expertise and experience in the treatment of the disease, and (3) the center has special equipment or facilities required for the study. As a result, the selection of centers in multicenter trials is not a random process but a much deliberated one. Fleiss (1986b) and Goldberg and Koury (1990) suggest that the center effects thus should not be considered random when performing statistical analyses. Note that the primary interest of a multicenter study is not for the comparison between centers. Hence, the centers should not be considered as a designed factor rather than a classification factor.

For assessment of the treatment effect in multicenter trials, it is suggested that a parameter be used for the overall treatment effect, which is the simple average over the treatment effects of J individual centers:

$$\begin{aligned} \delta &= \frac{1}{J} \sum_{j=1}^J \delta_j \\ &= \frac{1}{J} \sum_{j=1}^J (\mu_{Tj} - \mu_{Pj}). \end{aligned} \quad (12.5.9)$$

Denote \bar{d} as the simple average of d_j :

$$\begin{aligned} \bar{d} &= \frac{1}{J} \sum_{j=1}^J d_j \\ &= \frac{1}{J} \sum_{j=1}^J (\bar{Y}_{Tj} - \bar{Y}_{Pj}). \end{aligned} \quad (12.5.10)$$

Then \bar{d} is an unbiased estimator for δ with estimated variance

$$v(\bar{d}) = \frac{s^2 \sum w_j}{J^2}. \quad (12.5.11)$$

The null hypothesis that $H_0: \delta = 0$ is rejected with respect to a two-sided alternative at the α level of significance if

$$|t| = \left| \frac{\bar{d}}{s\sqrt{\sum w_j/J}} \right| > t(\alpha/2, \sum \sum (n_{ij} - 1)). \quad (12.5.12)$$

The $(1-\alpha)100\%$ confidence interval for δ is given by

$$\bar{d} \pm (s\sqrt{\sum w_j/J}) t(\alpha/2, \sum \sum (n_{ij} - 1)). \quad (12.5.13)$$

Note that an unbiased estimator for δ exists and that the test based on (12.5.12) is theoretically correct regardless of the sample sizes and presence of treatment-by-center interaction. When there is no interaction, the cell-means model (12.5.1) reduces to the main-effect model (12.5.2), which assumes that the treatment and center effects are additive. It follows that the treatment effect δ_j at center j is given by

$$\begin{aligned} \delta_j &= \mu_{Tj} - \mu_{Pj} \\ &= (\mu + \tau_T + \beta_j) - (\mu + \tau_P + \beta_j) \\ &= \tau_T - \tau_P, \quad j = 1, \dots, J. \end{aligned}$$

From the above it can be seen that due to the additivity of the main-effects model, the individual treatment effect within center j does not involve the center effects and is a constant across all centers. In addition, since δ in (12.5.9) is defined as the simple average over the treatment effects of J centers, under the main-effects model the overall treatment effect is also equal to $\tau_T - \tau_P$. Hence the overall treatment effect is a valid parameter for evaluation of treatment effects under both the cell-mean model and the main-effects model.

When there is no treatment-by-center interaction, the minimum variance unbiased estimator (MVUE) of $\tau_T - \tau_P$ can be obtained as

$$\bar{d}^* = \frac{\sum (d_j/w_j)}{\sum (1/w_j)}, \quad (12.5.14)$$

with the estimated variance given by

$$v(\bar{d}^*) = \frac{s^2}{\sum (1/w_j)}. \quad (12.5.15)$$

The hypothesis testing can be performed and confidence interval for $\tau_T - \tau_P$ can be constructed through the t statistic as described above.

The estimator for the overall treatment effect \bar{d}^* is the same as the combined estimator obtained from a post-treatment stratified analysis by considering each center as a stratum. The ratio $(1/w_j)/\sum(1/w_j)$ represents the information on the proportion of eligible patients in

the targeted population who are from center j . As a result, \bar{d}^* is still a consistent estimator in presence of treatment-by-center interaction which is close to the difference between the two treatment averages when the sample size is sufficiently large. Fleiss (1986b) points out that this ratio also reflects the ability and efficiency of patient recruitment for center j .

The null hypothesis of no treatment-by-center interaction is rejected at the α level of significance if

$$F_I = \frac{MS(AC)}{s^2} > F(\alpha, J - 1, \sum \sum (n_{ij} - 1)), \quad (12.5.16)$$

where $F(\alpha, J - 1, \sum \sum (n_{ij} - 1))$ is the α th upper quantile of a central F distribution with $J - 1$ and $\sum \sum (n_{ij} - 1)$ degrees of freedom. If the null hypothesis of no treatment-by-center interaction is rejected, it is helpful to determine whether the interaction is of quantitative or qualitative type. A quantitative interaction implies that the heterogeneity of the treatment effect across centers is due to the magnitude rather than the direction. On the other hand, a qualitative interaction indicates that the treatment effect is not only heterogeneous in magnitude but also changes direction from center to center as shown in Figure 2.4.5. Define

$$Q^- = \sum \frac{d_j}{s_j^2} I[d_j > 0]$$

and

$$Q^+ = \sum \frac{d_j}{s_j^2} I[d_j < 0] \quad (12.5.17)$$

Gail and Simon (1985) propose rejecting the null hypothesis of no qualitative interaction if

$$\min(Q^-, Q^+) > c, \quad (12.5.18)$$

where c is the critical value provided in Table 1 of Simon and Gail (1985).

In multicenter trials, although the sample size is selected to achieve a desired power for detection of the overall treatment effect, it is rarely large enough to identify the treatment-by-center interaction with adequate power. Fleiss (1986b) recommend the use of a 10% level of significance for the detection of treatment-by-center interaction. If the null hypothesis is rejected, then \bar{d} should be used for the inference of the overall treatment effect because it is an unbiased estimator for δ . Note that both d and d^* are unbiased for δ^* in the absence of treatment-by-center interaction. Since \bar{d}^* is the MVUE for δ^* , there are more degrees of freedom for estimation of σ^2 . Hence, if we fail to reject the null hypothesis of no treatment-by-center interaction, then δ^* and σ^2 should be estimated from the reduced main-effects model. However, if the number of centers is relatively small and the treatment-by-center interaction is not significant, both \bar{d} and \bar{d}^* will be close.

If all centers enroll very small numbers of patients, Goldberg and Koury (1990) suggest ignoring the centers in the analysis. For the unbalanced situations of patient enrollment among centers as presented in Tables 12.5.1 and 12.5.2, since no patients were randomized to treatment groups at some centers, the results and their interpretation are consistent if we ignore the center effect in the analysis. However, one should always follow the rule that if

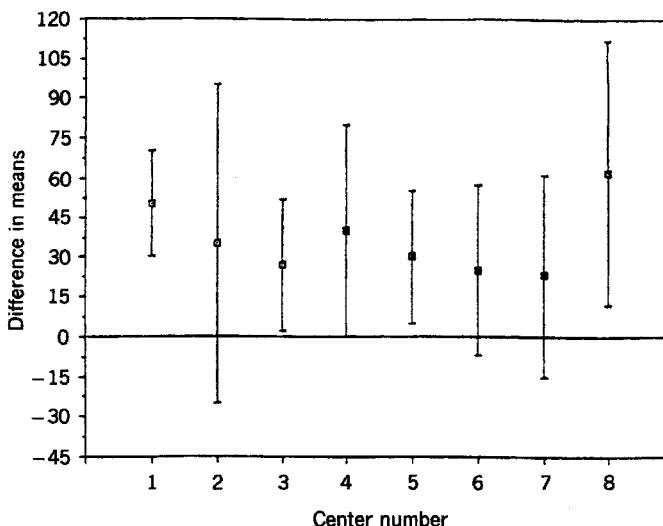


Figure 12.5.1 Difference in treatment means and 95% confidence intervals. Mean test drug change from the baseline minus the mean placebo change from the baseline. (Source: U.S. FDA Guideline for the Format and Content of the Clinical and Statistical Sections of an Application, 1988).

randomized then analyzed. One should never omit any centers from the analysis of multi-center trials. One way to resolve the situation, as illustrated in Table 12.5.1 and 12.5.2, is to randomly assign the patients at those centers to empty cells at other centers before the analysis of data.

Note that the overall treatment effect defined here is the simple average of center-specific treatment effects. Hence, if the qualitative treatment-by-effect interaction does exist, the overall treatment effect may be null. Although the statistical inference provided by \bar{d} is theoretically correct, both FDA and ICH guidelines require that not only individual center results be presented but also any extreme or opposite results among centers be noted and discussed. A graphical presentation of individual center results is provided in Figure 12.5.1. In addition all data, including demographic, baseline, post-baseline data, and efficacy data, should be presented by center.

12.6 MULTIPLICITY

Lepor et al. (1996) report the results of a double-blind, randomized multicenter clinical trial that evaluated the efficacy and safety of terazosin (10 mg daily), and α_1 -adrenergic-antagonist, finasteride, a 5 α -reductase inhibitor (5 mg daily) or both with a placebo control in equal allocation in 1229 men with benign prostatic hyperplasia. The primary efficacy endpoints of this trial are the American Urological Association (AUA) symptom score (Barry et al., 1992) and the maximum uroflow rate. These endpoints were evaluated twice during the four-week placebo run-in period and at 2, 4, 13, 26, 39, and 52 weeks of therapy. The primary comparisons of interest included pairwise comparisons among the active drugs and combination therapy, while the secondary comparisons consisted of a pairwise comparison of the active drugs and combination therapy with the placebo. The results for

the primary efficacy endpoints presented in Lepor et al. (1996) were obtained by performing analyses of covariance with repeated measurements based on the intention-to-treat population.

One of the objectives of the trial is to determine the time when the treatments reach therapeutic effects. Therefore comparisons among treatment groups were performed at each scheduled postrandomization visits at 2, 4, 13, 26, 39, and 52 weeks. In addition to the original observations of the primary endpoints, the change from baseline can also be employed to characterize the change after treatment for each patient. It may be of interest to see whether the treatment effects are homogeneous across race, age, and baseline disease severity. Therefore, some subgroup analyses can be performed such as for caucasians and for noncaucasians patients, for patients below or at least 65 years of age, for patients with the baseline AUA symptom score below 16 or at least 16, or for patients with the maximum uroflow rate below 10 or at least 10 ml/s. As a result, as illustrated in Table 12.6.1 the number of the total comparisons for the primary efficacy endpoints can be as large as 1344. If there is no difference among the four treatment groups and each of the 1344 comparisons are performed at the 5% level of significance, we can expect 67 statistically significant comparisons with reported *p*-values smaller than 0.05. The probability of observing at least one statistically significant difference among 1344 comparisons could be as large as 1 under the assumption that all 1344 comparisons are statistically independent. The number of *p*-values does not include those from the center-specific treatment comparisons and from other types of comparisons such as treatment-by-center interaction.

Although the above example is a bit exaggerated, it does point out that the multiplicity in multicenter clinical trials is an important issue that has an impact on statistical inference of the overall treatment effect. In practice, however, it is almost impossible to characterize a particular disease by a single efficacy measure due to (1) the multifaceted nature of the disease, (2) lack of understanding of the disease, and (3) lack of consensus on the characterization of the disease. Therefore, multiple efficacy endpoints are often considered to evaluate the effectiveness of test drugs in treatment of most diseases such as AIDS, asthma, benign prostatic hyperplasia, arthritis, postmenopausal osteoporosis, and ventricular tachycardia. Some of these endpoints are objective histological or physiological measurements such as the maximum uroflow rate for benign prostatic hyperplasia or pulmonary function FEV₁ (forced expiratory volume in one second) for asthma. Other may include

Table 12.6.1 Summary of Possible Number of Comparisons

Item	Number of Comparisons
Pairwise comparison	
Primary	3
Secondary	3
Visit	7
Primary end point	2
Response	2
Race	2
Baseline severity of disease	
AUA symptom score	2
Maximum uroflow rate	2
	1344

the symptoms or subjective judgment of the well-being of the patients improved by the treatments such as the AUA symptom scores for benign prostatic hyperplasia, asthma-specific symptom score for asthma, or the Greene climacteric scale for postmenopausal osteoporosis (Greene and Hart, 1987). Hence one type of multiplicity in statistical inference for clinical trials results from the source of multiple endpoints.

On the other hand, a clinical trial may be conducted to compare several drugs of different classes for the same indication. For example, the study by Lepor et al. (1996) compares two monotherapies of terazosin, finasteride with the combination therapy, and a placebo control for treatment of patients with benign prostatic hyperplasia. Some other trials might be intended for the investigation of a dose-response relationship of the test drug. For example, Gormley et al. (1992) evaluate the efficacy and safety of 1 and 5 mg of finasteride with a placebo control. This type of multiplicity is inherited from the fact that the number of treatment groups evaluated in a clinical trial is greater than 2. Other types of multiplicity are caused by subgroup analyses. Examples include a trial reported by the National Institute for Neurological Disorders and Stroke rt-PA Study Group (1995) in which stratified analyses were performed according to the time from the onset of stroke to the start of treatment (0–90 or 91–180 minutes). In addition, the BHAT (1982) and CAST (1989) studies were terminated early because of overwhelming evidence of either beneficial efficacy or serious safety concern before the scheduled conclusion of the trials by the technique of repeated interim analyses. In summary, multiplicity in clinical trials can be classified as repeated interim analyses, multiple comparisons, multiple endpoints, and subgroup analyses. Since the causes of these multiplicities are different, special attention must be paid to (1) the formulation of statistical hypotheses based on the objectives of the trial, (2) the proper control of experimentwise false positive rates in subsequent analyses of the data, and (3) the interpretation of the results. Since repeated interim analyses have been discussed in Chapter 10, in what follows, we will only address the remaining types of multiplicities.

Multiple Comparisons

Multiple comparisons are referred to as the comparisons among more than two treatments. If there is no structure among the treatment groups, then Bonferroni's technique discussed in Section 8.4 is appropriate. The concept of Bonferroni's technique is to adjust the *p*-values for control of experimentwise type I error rate α for pairwise comparisons. The method is useful when the number of treatments is small. In addition Bonferroni's method does not require that the structure of the correlation among comparisons be specified nor that the number of patients in each treatment group be equal. However, when the number of treatment groups increases, Bonferroni's adjustment for *p*-values becomes very conservative and may lack adequate power for the alternative in which most or all efficacy endpoints are improved. To overcome this drawback, many modified Bonferroni procedures have been proposed. For illustration purposes, in what follows we will only introduce modified procedures proposed by Holm (1979) and Hochberg (1988).

Holm (1979) proposes that Bonferroni's procedure be modified as follows: Let $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$ be the ordered values of *p*-values p_1, p_2, \dots, p_m obtained from testing the following hypotheses:

$$\begin{aligned} H_{k0}: \mu_i - \mu_j &= 0 \\ \text{vs. } H_{ka}: \mu_i - \mu_j &\neq 0, i \neq j; k = 1, \dots, m. \end{aligned} \quad (12.6.1)$$

Holm's procedure starts with the ordered hypothesis $H_{(1)0}$, which is the hypothesis with the smallest p -value, $p_{(1)}$:

- Step 1. Stop the procedure and conclude that all pairwise comparisons are not statistically significant at the α level of significance if $mp_{(1)} > \alpha$. Otherwise, reject the null hypothesis $H_{(1)0}$ and continue to step 2.
- Step 2. Stop the procedure and conclude that the remaining $(m - 1)$ pairwise comparisons are not statistically significant at the α level of significance if $(m - 1)p_{(2)} > \alpha$. Otherwise, reject the null hypothesis $H_{(2)0}$ and continue to the next step.
- \vdots
- Step k . Stop the procedure and conclude that the rest of the $(m - k + 1)$ pairwise comparisons are not statistically significant at the α level of significance if $(m - k + 1)p_{(k)} > \alpha$. Otherwise, reject the null hypothesis $H_{(k)0}$ and continue to the next step. Repeat step k until it stops.

The Holm's Bonferroni's procedure is the same as the original Bonferroni's procedure for testing $H_{(1)0}$ at which the p -value is adjusted down by dividing the nominal level of significance by the number of total comparisons. But for the rest of $(m - 1)$ hypotheses, Holm's procedure is much sharper than Bonferroni's procedure, since it only requires that the p -values be adjusted according to the number of remaining pairwise comparisons. Because Holm's procedure requires the smallest p -values to be smaller than α/m , Hochberg (1988) suggests another version of Bonferroni's procedure which is even sharper than Holm's procedure. Hochberg's procedure begins with the ordered hypothesis $H_{(m)0}$ corresponding to the largest p -value $p_{(m)}$.

- Step 1. Stop the procedure and conclude that all m pairwise comparisons are statistically significant at the α level of significance if $p_{(m)} \leq \alpha$. Otherwise, do not reject the null hypothesis of $H_{(m)0}$ and continue to step 2.
- Step 2. Stop the procedure and conclude that the remaining $m - 1$ pairwise comparisons are statistically significant at the α level of significance if $2p_{(m-1)} \leq \alpha$. Otherwise, do not reject the null hypothesis of $H_{(m-1)0}$ and continue to the next step.
- \vdots
- Step k . Stop the procedure and conclude that all $(m - k + 1)$ pairwise comparisons are statistically significant at the α level of significance if $(m - k + 1)p_{(m-k+1)} \leq \alpha$. Otherwise, do not reject the null hypothesis of $H_{(m-k+1)0}$ and continue to the next step. Repeat step k until it stops.

Hochberg's procedure is more powerful than Bonferroni's procedure or Holm's procedure because it only requires that the largest p -value be smaller than α to declare one statistically significant comparison.

Sometimes, clinical trials are planned to compare the effectiveness and safety of K doses of a test drug versus a placebo group to determine the minimum effective and maximum tolerable dose. As a result, in addition to characterization of the dose-response relationship, pairwise comparisons between each dose and the placebo are of clinical interest. In this case the hypothesis of interest can be formulated as follows:

$$\begin{aligned} H_{k0}: \mu_i - \mu_0 &= 0, \\ \text{vs. } H_{ka}: \mu_i - \mu_0 &\neq 0, \quad i = 1, \dots, K. \end{aligned} \tag{12.6.2}$$

Dunnett's procedure for comparison with a control can be directly applied (Dunnett, 1955). Let $\bar{Y}_i - \bar{Y}_0$ be the observed sample mean difference between dose i and the placebo, and let $v(\bar{Y}_i - \bar{Y}_0)$ denote the estimated variance of $\bar{Y}_i - \bar{Y}_0$, $i = 1, \dots, K$. Then hypothesis (12.6.2) is rejected at the α level of significance if the $(1 - \alpha)100\%$ simultaneous confidence interval for $\mu_i - \mu_0$,

$$\bar{Y}_i - \bar{Y}_0 \pm \sqrt{v(\bar{Y}_i - \bar{Y}_0)} t(\alpha, K, dfE, \rho_{ij}) \quad (12.6.3)$$

does not contain zero, where $t(\alpha, K, dfE, \rho_{ij})$ is the critical value of the two-sided comparison for comparing K treatments with a control as given in Hochberg and Tamhane (1987), dfE is the error degrees of freedom from the appropriate analysis of variance table, and ρ_{ij} is the correlation between $\bar{Y}_i - \bar{Y}_0$ and $\bar{Y}_j - \bar{Y}_0$ which is given by

$$\rho_{ij} = \sqrt{\frac{n_i n_j}{(n_i + n_K)(n_j + n_K)}}.$$

Dunnett's procedure for the one-sided alternative is also available (Dunnett, 1955). In addition step-down and step-up versions of Dunnett's procedure for comparing treatments with a control are proposed by Dunnett and Tamhane (1991, 1992) for various multiple comparisons of treatments with a control. An overview of multiple comparisons in clinical trials can be found in Dunnett and Goldsmith (1995).

Multiple Endpoints

As mentioned earlier, the efficacy of a test drug in treatment of a certain disease can be characterized through multiple clinical endpoints. Capizzi and Zhang (1996) classify the clinical endpoints into primary, secondary, and tertiary categories whose criteria are modified and given below:

Primary endpoints should satisfy the following criteria:

- Should be of biological and/or clinical importance.
- Should form the basis of the objectives of the trial.
- Should not be highly correlated.
- Should have sufficient power for the statistical hypotheses formulated from the objectives of the trial.
- Should be relatively few (e.g., at most 4).

For secondary endpoints, typical criteria include whether the endpoints are (1) biologically and/or clinically important, but with less adequate statistical power, (2) potentially important, but highly correlated with primary endpoints, and (3) address other important, but ancillary objectives. The criteria of tertiary endpoints depend on whether (1) they are exploratory endpoints and (2) they are not of major importance.

Since the sample size of a clinical trial is usually selected to provide a sufficient power for detection of a difference in some or all primary clinical endpoints, we focus on the issue of false positive and false negative rates caused by multiple primary endpoints. It is, however, very important to understand the objective of the trial and to tailor and formulate the corresponding statistical hypotheses to the specific objectives in terms of multiple endpoints. Consider a randomized, parallel group, double-blind trial that evaluates the efficacy

of a test drug versus a placebo control through a set of K primary endpoints. Let Y_{ijk} denote the k th endpoint for the j th subject in treatment group i ; $k = 1, \dots, K$, $j = 1, \dots, n_i$, $i = T, P$. Define the population average for the k th endpoints of group i as

$$E(Y_{ijk}) = \mu_{ik}, \quad k = 1, \dots, K, i = T, P.$$

The following statistical hypothesis formulates the clinical objectives of the trial. We declare that the test drug is efficacious if the test drug demonstrates a superior efficacy in any one of the K primary endpoints under the assumption that large values are better.

$$\begin{aligned} H_0: \mu_{Tk} &= \mu_{Pk} && \text{for all } k, \\ \text{vs. } H_a: \mu_{Tk} &> \mu_{Pk} && \text{for at least one } k = 1, \dots, K. \end{aligned} \quad (12.6.4)$$

Let $\bar{d}_k = \bar{Y}_{Tk} - \bar{Y}_{Pk}$ be the observed difference in sample mean of endpoint k between the test drug and the placebo. Also let $v(\bar{Y}_{Tk} - \bar{Y}_{Pk})$ be the estimated variance of $\bar{Y}_{Tk} - \bar{Y}_{Pk}$, $k = 1, \dots, K$. Denote p_k as the observed p -value based on the test statistics corresponding to hypothesis (12.6.4):

$$t_k = \frac{\bar{Y}_{Tk} - \bar{Y}_{Pk}}{\sqrt{v(\bar{Y}_{Tk} - \bar{Y}_{Pk})}}, \quad k = 1, \dots, K. \quad (12.6.5)$$

Then Bonferroni's adjustment of the p -value described above can be directly applied to control the false positive rate based on p_1, \dots, p_K . Table 12.6.2 provides the Bonferroni correction and the true nominal significance level to maintain the overall false positive rate of 5% for various correlation coefficient when the number of endpoints is 4. In general, if the number of primary endpoints is small and the correlation between them is less than 0.5, Bonferroni's correction of p -value works reasonably well. However, when the number of endpoints increases and their correlations become large, then Bonferroni's adjustment will be too conservative and lack the power to detect any treatment effects in the primary endpoints. The criteria for primary endpoints specify that the number of primary endpoints should be at most 4 and that the correlation should not exceed 0.5. Both Bonferroni's technique and its modification proposed by Hochberg (1988) are appropriate methods for adjustment of p -values for hypotheses of (12.6.4).

Note that Bonferroni's method achieves its greatest power when the true treatment effect exists in only one of the K endpoints. However, some early phase II trials may have a large number of clinical endpoints that are highly correlated one another. In phase III

Table 12.6.2 Bonferroni's Correction and the True Nominal Significance Level to Maintain the Overall False Positive Rate of 5% for Various Correlation Coefficients the Number of End Points is 4

Bonferroni's Correction	Correlation Coefficient										
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
α level (%)											
	1.25	1.27	1.30	1.38	1.40	1.54	1.70	1.84	2.20	2.69	5.0

studies it is useful to obtain a single composite index from multiple endpoints that provides a summary of the treatment's efficacy. Several methods for the construction of a composite index have been proposed by O'Brien (1984), Pocock et al. (1987), and Tang et al. (1989).

Assume that Y_{ijk} defined above have been standardized (i.e., subtracting the overall mean from each observation and dividing by the within-group sample standard deviation). Let $\bar{\mathbf{d}}$ be the vector of the observed sample mean differences of the K endpoints and \mathbf{S} denotes the estimated within-group correlation matrix obtained under the assumption of the common covariance matrix of the K endpoints for both treatment groups. Then the global test statistic under the normality assumption proposed by O'Brien (1984) is an estimated generalized least squares (EGLS) method, which is given by

$$T_c = \frac{\mathbf{1}'\mathbf{S}^{-1}\bar{\mathbf{d}}}{\sqrt{w(\mathbf{1}'\mathbf{S}^{-1}\mathbf{1})}} \quad (12.6.6)$$

where $\mathbf{1}$ is the $k \times 1$ vector of 1 and $w = (1/n_T) + (1/n_P)$. The null hypothesis is rejected at the α level of significance if

$$T_c > t(\alpha, N - 2K),$$

where $t(\alpha, N - 2K)$ is the α th upper quantile of a central t distribution with $N - 2K$ degrees of freedom. The numerator of the global test statistic in T_c is a linear combination of the K observed sample mean differences with weights equal to the column sums of \mathbf{S}^{-1} . As a result the endpoints that are highly correlated with other endpoints receive small weights.

When the normality assumption for the distribution of multiple endpoints is in doubt, a nonparametric procedure (O'Brien, 1984) is also available for hypothesis (12.6.4). The nonparametric method is a rank-sum type test that starts ranking Y_{ijk} among all observations of endpoints k in the combined sample. Let us denote R_{ij} as the sum of the ranks assigned to the j th patient in group i , $j = 1, \dots, n_i$, $i = T, P$. We can then apply the Wilcoxon rank-sum test to R_{ij} for testing hypothesis (12.6.4).

Pocock et al. (1987) extends O'Brien's procedure to combine multiple binary clinical endpoints which are commonly used to evaluate the efficacy of patients' response to the treatment. The test statistic of their procedure for combining multiple binary endpoints is to replace the elements in vector $\bar{\mathbf{d}}$ and the estimated correlation matrix \mathbf{S} by $p_{Tk} - p_{Pk}$ and $s_{kk'}$, $1 \leq k \neq k' \leq K$, respectively, where p_{Tk} and p_{Pk} are the proportions of patients responding to the test and placebo groups, respectively, and

$$s_{kk'} = \frac{p_{kk'} - p_k p_{k'}}{\sqrt{p_k p_{k'}(1-p_k)(1-p_{k'})}}, \quad (12.6.7)$$

where $p_{kk'}$ is the proportion of patients with responses in both endpoints k and k' , and p_k are the proportion of the patients in both groups whose response was defined in (9.3.10). If the sample size is moderately large, then the global test statistic T_c follows a standard normal distribution.

Pocock et al. (1987) propose an extension of the O'Brien EGLS method that combines censored and binary endpoints as are common in cancer chemotherapy trials. For example, let endpoint k be the censored survival time and endpoint k' be the tumor response. The test statistic is obtained by substituting the elements of $\bar{\mathbf{d}}$ and \mathbf{S} in test statistic T_c by z_k and $s_{kk'}$, respectively, where z_k is the square root of the logrank statistic X_{LR} defined in (10.3.5), and

$$s_{kk'} = \frac{R_d - \sum p(t)}{N \sqrt{v_k p_k (1-p_k)}}, \quad (12.6.8)$$

where R_d is the total number of the patients who responded to the treatment but nevertheless died, v_k given by the denominator of the logrank statistics in (10.3.5), p_k is defined above in (12.6.7), $p(t)$ is the proportion of patients who responded among those who were at risk at time t , and the summation is over all death times.

The null hypothesis (12.6.4) assumes that all K endpoints are equally important. As a result the weights for the test statistics based on the EGLS method given above are only based on the pairwise correlations between endpoints. However, if the relative importance of endpoints can be specified a priori in the protocol with known weights c_k , $k = 1, \dots, K$, then the global test statistic given in (12.6.6) can be easily modified to accommodate the unequal priorities among the K endpoints as follows:

$$T_c = \frac{\mathbf{1}'(\mathbf{C}\mathbf{S}\mathbf{C})^{-1}\bar{\mathbf{d}}}{\sqrt{w[\mathbf{1}'(\mathbf{C}\mathbf{S}\mathbf{C})^{-1}\mathbf{1}]}} \quad (12.6.9)$$

where \mathbf{C} is a $k \times k$ diagonal matrix with diagonal elements being c_k .

The objective of clinical trials for a certain class of drug products with respect to some diseases may be more stringent than that for the formulation of hypothesis (12.6.4), and the test drug may be claimed to be efficacious only if it demonstrates a statistically significantly superior effect in all of the K primary endpoints. In other words, the corresponding statistical hypothesis is then expressed as

$$\begin{aligned} H_0: \mu_{Tk} &= \mu_{Pk} && \text{for at least one } k, \\ \text{vs.} \quad H_a: \mu_{Tk} &> \mu_{Pk} && \text{for all } k, k = 1, \dots, K. \end{aligned} \quad (12.6.10)$$

The null hypothesis (12.6.4) is that the treatment difference between the test drug and placebo in all K endpoints are zero versus the alternative that the test drug is superior in at least one of the K endpoints. On the other hand, the null hypothesis (12.6.10) indicates that the treatment difference between the test drug and placebo is zero in any one of the K endpoints while the alternative hypothesis of (12.6.10) requires that the test drug be effective in all K endpoints. Hence, the intersection-union test (IUT) proposed by Berger (1982) can be applied to test hypothesis (12.6.10) which is rejected at the α level of significance if and only if each of the K individual null hypotheses H_{k0} is rejected at the α significance level, $k = 1, \dots, K$. Although the size of the IUT procedure is α (i.e., an α size test), it is very conservative as demonstrated in Table 12.6.3. Using a simulation, Capizzi and Zhang (1996) provide the nominal significance level at which each individual hypothesis is tested to achieve an experimentwise false positive rate of 5%. The simulation result is reproduced in Table 12.6.4 for four endpoints. For example, if the correlation among four endpoints is equal to 0.4, then for each endpoint the hypothesis must be tested at a nominal level of 0.289 to maintain an experimentwise type I error rate of 0.05. However, this approach with

Table 12.6.3 False Positive Rate (%) for the IUT Procedure Performed at the Significance Level When the Number of End Points is 4

Correlation Coefficient					
0	0.2	0.4	0.6	0.8	1.0
<0.01	0.02	0.14	0.5	1.3	5

Source: Capizzi and Zhang (1996).

Table 12.6.4 Nominal Significance Level (%) for the IUT Procedure to Maintain an Experimentwise False Positive Rate of 5% when the Number of Endpoints is 4

Correlation Coefficient					
0	0.2	0.4	0.6	0.8	1.0
47.3	37.6	28.9	20.9	13.6	5

Source: Capizzi and Zhang (1996).

a nominal level 0.289 for each of four tests does not have a size of 5%. Suppose that there is a ranking among the K endpoints. The individual hypotheses for the important primary endpoints are tested at the 5% level as usual. However, for other endpoints with less relevance, individual tests can be performed at a higher nominal significance level to improve the power of the IUT procedure. But the resulting IUT procedure does not have an experimentwise false positive rate of 5%. In general, if the hypothesis for each of the K endpoints is tested at α_k , $k = 1, \dots, K$, then the experimentwise false positive rate (i.e., size) of the IUT procedure for hypothesis (12.6.10) is the maximum of $\alpha_1, \dots, \alpha_K$ (Berger and Hsu, 1996).

Hypothesis (12.6.4) is somewhat too liberal in the sense that the declaration of the test drug's effectiveness only requires demonstration of superior efficacy at one of the K endpoints. In contrast is the strict requirement of superior efficacy at all K endpoints demanded by hypothesis (12.6.10). As a result Capizzi and Zhang (1996) propose a hybrid hypothesis in which the test drug is claimed to be effective if superior efficacy is demonstrated at M of K endpoints. The corresponding statistical hypotheses are given as

$$\begin{aligned} H_0: \mu_{Tk} &= \mu_{Pk} && \text{for at most } K-M \text{ endpoints,} \\ \text{vs. } H_a: \mu_{Tk} &> \mu_{Pk} && \text{for at least } M \text{ of } K \text{ endpoints,} \end{aligned} \quad (12.6.11)$$

$k = 1, \dots, K$. Note that hypothesis (12.6.10), which requires a demonstration of superior efficacy at each of the primary endpoints, is often considered the hypothesis of choice for declaring the effectiveness of a drug. For example, finasteride at 5 mg was approved by U.S. FDA in treatment of patients with benign prostatic hyperplasia because its effectiveness was demonstrated at all three primary endpoints such as a statistically significant decrease in total urinary-symptom scores (p -value < 0.001), a statistically significant increase of 1.6 ml/s in maximal urinary-flow rate (p -value < 0.001), and a statistically significantly 19% reduction in prostatic volume (p -value < 0.001). On the other hand, hypotheses (12.6.4) and (12.6.11) may be appropriate for clinical trials during phase II development because the primary endpoints may not be fully understood for some diseases or the effects of the drug on different aspects of the disease are still under investigation.

Subgroup Analysis

Although the primary objective of clinical trials is to provide a valid and unbiased inference of the treatment effect for the disease under study, it is of great interest to observe whether the treatment effect is consistent across some demographic factors such as age, gender, race, baseline disease severity, some prognostic factors, or previous medical conditions and concomitant medications. As a result, subgroup analyses are performed within

strata according to some stratification or classification factors. The primary goal of these subgroup analyses is not for a definitive statistical inference of the treatment effect for each subgroup but rather for an exploratory identification of unusual or unexpected results.

The difference between subgroup analyses and multiple comparisons and multiple endpoints is that patients in subgroups stratified by different values of a covariate are different and the test statistics obtained from different subgroups are statistically independent. For example, consider a randomized double-blind, parallel-group clinical trial for the investigation of a new antihypertensive agent's effectiveness versus a placebo. The primary endpoint is the change from the baseline in diastolic blood pressure (mmHg). Suppose that it is of interest to investigate whether the reduction in diastolic blood pressure is consistent across age groups. Therefore, patients are divided into groups according to age groups of 18–30, 31–50, 51–70, and over 70. The two-sample unpaired t -statistic can be computed within each age group to test whether the new antihypertensive agent is efficacious in each of the four age groups. Since patients in each age group are different and these age subgroups are mutually exclusive, the four unpaired t -statistics are independent of each other. As a result the adjustment of p -values is straightforward. Let α be the desired experimentwise false positive rate for a total of K subgroups. Then, the nominal significance level α^* of the test performed for each subgroup is given as

$$\alpha^* = 1 - (1 - \alpha)^{1/K} \quad (12.6.12)$$

Since the purpose of subgroup analyses includes implicit future hypotheses, the experimentwise false positive rate must be adequately controlled and conservative methods such as Bonferroni's correction of p -values are often used. However, the ICH guidelines on structure and contents of clinical study reports require that *where there is a prior hypothesis of a differential effect in a particular subgroup, this hypothesis and its assessment should be part of the planned statistical analysis*. Even so, it is recommended that the clinical reports should state that their subgroup analyses are exploratory. On the other hand, CPMP also issues a note for *Guidance on Biostatistical Methodology in Clinical Trials in Applications for Marketing Authorizations for Medicinal Products* which expresses a different view on subgroup analyses (CPMP, 1995). The note emphasizes that the model approaches such as analysis of covariance, logistic regression, or Cox's proportional hazard model should include covariates and treatment-by-covariate interactions in order to obtain an overall treatment effect. The note does not recommend the multiple separate analyses within strata defined by the covariates. Under the assumption of no treatment-by-covariate interaction, d^* given in (12.5.14) is the minimum variance unbiased estimator for the overall treatment effect.

12.7 DATA MONITORING

Definition and Objectives

Data monitoring is an active process involving a completely blinded review of clinical data while a trial is in progress (PMA, 1993). O'Neill (1993) provides some viewpoints from a regulatory perspective. He suggests that data monitoring should include the following:

1. Assessments of the quality and relevance of the data and the extent to which the protocol is being followed.
2. Arrangements for data processing, auditing, cleaning, and report generation.

3. Compilation and an assessment of safety data.
4. Compilation and an assessment of efficacy data.
5. Points in time for decision making on whether to continue or stop a trial based on observed results.

Data monitoring is multifaceted process that can vary from stage to stage during the clinical development of a drug product. For example, during phases I and II clinical studies are mainly of an exploratory nature, whereas during phase III they take a confirmatory form. In addition, the purpose might differ depending on the sponsors such as the pharmaceutical industry and the U.S. National Institutes of Health. However, if a clinical trial demonstrates overwhelming evidence of beneficial efficacy or unexpected harmful adverse effects for the test drug, then it is recommended that the trial be terminated before the scheduled completion of the trial. Data monitoring and interim analyses are commonly employed for clinical trials on the treatment of life-threatening diseases or severely debilitating illnesses with long-term follow-up and endpoints such as mortality or irreversible morbidity. The decision for early termination of such clinical trials is crucial and cannot be made simply based on an application of one of many statistical procedures for interim analyses discussed in Chapter 9. Data monitoring and interim analysis are extremely complex processes for which there exist no definitive rules, and they are critical to good clinical practice. Canner (1981) points out in decision making in clinical trials, no single statistical decision rule or procedure can take the place of well-reasoned considerations of all data by a group of concerned, competent, and experienced persons with a wide range of scientific backgrounds and points of view. In particular, the decision for early termination of a clinical trial should be based on such considerations as baseline comparability, unbiased evaluation, compliance, internal consistency with other endpoints and subgroup analyses, external consistency, benefit and risk, length of follow-up, public impact, repeating testing, and multiple comparisons.

Regulatory Concerns

Improper and sloppy data monitoring and interim analyses can potentially alter the conduct of a clinical trial. Consequently, serious bias may be introduced in the treatment effects. Both the FDA (FDA, 1988) and ICH guidelines (ICH, 1996) on the format, structure, and contents of a clinical report explicitly state that:

The process of examining and analyzing data accumulating in a clinical trial, either formally or informally, can introduce bias and/or increase type I error. Therefore, all interim analyses, formal or informal, pre-planned or ad hoc, by any study participant, sponsor staff member, or data monitoring group should be described in full, even if the treatment groups were not identified. The need for statistical adjustment because of such analyses should be addressed. Any operating instructions or procedures used for such analyses should be described. The minutes of meetings of any data monitoring group and any data reports reviewed at those meetings, particularly a meeting that led to a change in the protocol or early termination of the study, may be helpful and should be provided. Data monitoring without code-breaking should be also described, even if this kind of monitoring is considered to cause no increase in type I error.

In addition, as indicated by O'Neill (1993), the FDA classifies the issues of data monitoring in clinical trials as planning, reporting, operation, and documentation of the trials. These issues include (1) unreported interim analyses, (2) planned or unplanned interim access to unblinded comparative study results, (3) failure of assessment of impact of unplanned

interim analyses on study results, (4) bias on the future conduct of the trial caused by unblinded assess to study results, (5) the recognition of all relevant parties of the regulatory implications of early termination of trials, (6) development of efficient, effective communication and information flow between the data-monitoring committee and the regulatory authority, (7) appropriate evaluation of exploratory trials, (8) planning trials not to stop early for efficacy reasons alone but to balance the need for safety data on longer-term exposure with short-term follow-up of early efficacy results, (9) establishment of policies regarding access to ongoing data, access to unblinded data, and participation in the decision-making chain.

The above regulatory issues on data monitoring and interim analyses are concerned with a potential bias in estimating the treatment effect, an inflation of a false positive rate, and documentation of the processes. Since data monitoring and interim analyses can affect the subsequent conduct of the trial, a bias may be introduced that may not be measurable or quantifiable. As a result the generalizability of the trial results (or inference) is placed in serious jeopardy. In addition the integrity and credibility of the trial becomes seriously affected. Therefore it is important to make any attempts/efforts to eliminate both known and unknown biases. As discussed earlier in this book, one way to avoid a potential bias is to maintain blindness throughout the study. To achieve this objective, standard operating procedures should be followed in the selection of plans, methods, data management, documentation of data monitoring, planned and unplanned interim analyses, and dissemination of the results.

In general, phase IIB, phase III, and phase IV trials must be triple-blinded in the sense that patients, investigators, and the clinical project team responsible for the development are unaware of any individual treatment assignments and treatment results until all data of the trial have been received, entered, edited, and verified and the database locked for analysis. According to the definition given by the PMA, data monitoring is to be used for all trials with and without planned interim analyses. In principal, all parties except for the project statistician, data coordinator, and programmer are kept blinded until after the database is locked and the statistical analyses completed, reviewed, and approved. When an interim analysis is required, then only a statistician who is not involved in the trial is authorized to merge the file of randomization codes of the individual treatment assignments with the dataset of endpoints by patient number required for the interim analysis. The file of randomization codes must be kept in a classified and secure area in a database management system that is separate from the rest of the trial data and can only be accessed by the authorized personnel. The extent of data included in the dataset for the interim analysis should be (1) specified in the protocol and (2) minimal in order to satisfy the objectives of the interim analysis and information for any subsequent decision-making process leading to termination of the trial. On the other hand, the results of the interim analysis could be presented as an overall summary with proper treatment identification for a meaningful interpretation but excluding individual patient listings. These and any results of interim analyses should be accessible only to a few designated people such as the members of an independent data monitoring committee. A standard operating procedure must be established to maintain the blindness of the study.

Early and Late Stages of Drug Development

As was mentioned earlier, the objectives of data monitoring depend on the stage of drug development. In the early stages the trials are often exploratory in nature. As a result the objectives of data monitoring and interim analyses are in the context of trial management such as verification of design assumptions, detection of possible design flaws, monitoring

for unexpected side effects, and generation of hypotheses for future use. Although general principles of data monitoring can be applied to clinical trials at early phases of drug development, their protocols usually do not specify the objectives, procedures, frequencies, and methods of interim analyses. In addition, since these trials always proceed to completion regardless of any interim analyses, the *p*-values obtained from interim analyses remain generally unadjusted. Consequently they should be considered as descriptive rather than inferential. Moreover, since clinical trials at an early stage are focused on generating hypotheses to be confirmed at later stages, decisions made for future trials are sometimes based on incomplete information. Accordingly, these decisions may be misleading due to (1) unadjusted *p*-values of interim analyses and (2) small sample sizes of these studies. It is suggested that data monitoring and interim analyses for these early trials be planned with respect to dates of interim analyses, numbers of patients, summary statistics, and any other adequately documented decisions.

Clinical trials conducted in the late stage of clinical development are confirmatory, adequate well-controlled studies to provide substantial evidence of effectiveness and safety for approval. On the other hand, phase IV studies and mega trials such as physician health study, ISIS series trials, or GUSTO study are conducted to evaluate the drugs or compare different agents in various combinations for the same indication in a much larger targeted population. These trials are usually very large and involve more patients and take much longer to complete with huge human and financial investment. Therefore, for ethical and scientific reasons interim analyses are planned for these trials, in particular, with survival and irreversible morbidity such as stroke or myocardial infarction as primary endpoints. However, early termination of a trial conducted in non-life-threatening diseases solely due to efficacy is not common. The guidelines for data monitoring and interim analyses for large confirmatory trials include:

1. Group sequential methods introduced in Chapter 10 should be described in the protocol with respect to the required primary endpoints, frequency, boundaries, and references of the employed methods for interim analysis and decision rules derived partly from the results of interim analyses.
2. An independent data monitoring committee with internal or external members should be established to monitor the data and to review the results of interim analyses.
3. Principles of adequate documentation with respect to the results of interim analyses and minutes of data monitoring committee are essential for good practice of data monitoring.
4. Blindness on dissemination of the results of interim analyses and deliberation of data monitoring committee should be strictly enforced and tightly controlled.

Administrative Interim Analyses

From a functional point of view, interim analyses by the data-monitoring committees can be classified as *formal* and *administrative analyses* (PMA, 1993; Williams et al., 1993). The aim of a formal interim analysis is to decide on early termination of a planned study if there is compelling evidence of beneficial effectiveness or harmful side effects. The administrative interim analysis is relevant in pharmaceutical cases and is conducted for the reasons external to a trial at the request of regulatory agencies or upper management. Such an administrative analysis rarely concerns early termination, and it is usually performed at an

early stage of the trial in order to verify the design assumptions, check the data entry, editing, and review processes, and compare the simple baseline summary statistics used in the enrollment of patients. Nevertheless, sometimes trials are terminated this early due to major design flaws, an unusual placebo effect or unexpected toxicity, and other reasons relating to recruitment problems, budget considerations, administrative cutoffs, company merges, and contracting the study out to a contract research organization (CRO). For these reasons standard operating procedures for early termination based on an administrative analysis must be established. The standard operating procedures should identify the individual responsible for authorizing an administrative interim analysis and outline a clear paper trail for the reasons and actions as a consequence of such an analysis. It is important to recognize that the potential danger of any administrative interim analysis lies in introduction of bias and inconclusive evidence of trend because of lack of power.

Data Monitoring Committee

An independent data monitoring committee (DMC) should be established for any confirmatory trials with planned interim analysis, in particular, for the trials conducted in life-threatening diseases or severely debilitating ailments. The model of data monitoring committee formulated by the U.S. National Institutes of Health (NIH model) is usually adopted for the trials sponsored by government. The model of the NIH clinical trials consists of sponsor, steering committee, center or principal investigators, data coordination and statistical analysis center, central laboratories, and data monitoring committee. The members of the NIH DMC may include those with the disciplines in clinical, laboratory, epidemiology, biostatistics, data management and ethics. In order to be independent, the members and their family should have no conflict of interest, such as, no financial holdings in companies. In addition, the members should not discuss nor disseminate the results of the interim analyses outside DMC. A member representing the sponsor is usually included in DMC as a nonvoting member and is allowed to attend the open session of DMC meetings only.

The responsibilities of DMC are to monitor the safety and ethical aspects of the trial with respect to the patients, investigators, sponsors, and the regulatory authorities in descending order of priority. DMC achieves these responsibilities by review of protocols; interim review of study progress and the quality of trial conduct, monitoring safety data and possible efficacy and benefit, and recommendation of early termination of trials and dissemination of the primary results. Documentation is also a crucial part of functions of DMC which include protocol, operational manuals for key data, decisions processes, and different situations that might be encountered during the conduct of the trial and their possible resolutions, and interim data reports and minutes of the meetings.

The DMC should be prepared and ready from the start and should concentrate on the primary efficacy and safety endpoint but not on individual case reports. Therefore, an on-line data management and analysis system is essential for minimization of delay in data entry and event verification so that any decision made by DMC is on the currently available data. As a result, sometimes, an independent data coordination and statistical analysis center is usually established for trials sponsored by government. The data coordination and statistical analysis center should be independent of the sponsor and responsible for case report design, on-line data management system for data entry, editing and verification, and for the quality control of the conduct of the trial through training, certification, tracking of case report forms and reports, and design and maintenance of on-line analysis system for interim and final analyses. In order to provide a good practice of data monitoring and to

work problematically, the data coordination and statistical analysis center is required to function and interact closely with the sponsor, clinics, steering committee, and most importantly, the data monitoring committee.

The philosophy for decision-making process taken by DMC should be *be ahead of time*, *be prepared*, and *be problematic*. Various formats of DMC meetings can be set up for different objectives. The first one is an open session in which the representative from the sponsor might attend. The objective of the open session is to blindly review the progress, conduct, recruiting, and logistic issues of the trial. The second format is a closed session at which the members of DMC review the summary efficacy and safety results from interim analyses by treatment group. The real treatments or dummy treatment codes such as A versus B could be used for review of interim results. However, the sponsor is not allowed to attend the close session. The last format is the executive session at which the decision of early termination will be deliberated and made. Sometimes, a quorum is set up for any decision about trial termination. This quorum usually include the statistician of DMC.

The DMC might report to different parties such as the sponsor, the study chair, or the executive or steering committee depending on the model and organization of the trial. For example, the Cooperative North Scandinavian Enalapril Survival Study II (CONSENSUS II) was a Scandinavian study sponsored by a U.S. pharmaceutical company with a planned sample size of 4500 patients to compare enalapril, an angiotensin converting enzyme (ACE) inhibitor, to a placebo in the treatment of acute myocardial infarction (Williams et al., 1993). This trial had a steering committee of 12 members, including a nonvoting sponsor's representative and a DMC consisting of three clinicians and one statistician from the academe and a nonvoting sponsor's statistician. The responsibility of the DMC for this trial was to review unblinded analyses and to make recommendations to the steering committee. The sponsor's statistician at the DMC, who was the unblinded sponsor's employee for this trial, provided analyses to the DMC.

As was mentioned earlier, interim analyses are an established part of the pivotal phase III trials conducted by pharmaceutical companies with mortality or irreversible morbidity as the primary endpoints. The method of the DMC can vary according to external involvement of the trial. For certain trials with non-life-threatening diseases, a complete in-house DMC may be set up for data monitoring and interim analyses. The next level of data monitoring and interim analyses is an independent external DMC but the analyses must be performed blinded internally, such as in CONSENSUS II. Sometimes a trial may have an external DMC with data processing performed internally but with the interim analyses done externally. The sponsor of the trial is allowed to attend the open sessions of the DMC meetings. This structure and function is well accepted. Under this structure the sponsor must have sophisticated computer and data management systems and sufficient resources to process a large volume of data. Data are monitored blindly and independently, and interim analyses are performed independently too. However, an independent external DMC may have both data processing and analyses performed externally. The most extreme method of the DMC is a *complete hands-off* approach with the external data processing and analyses done with no sponsor's representative. A typical example is the GUSTO trial.

Regulatory authorities such as the FDA and ICH have expressed their interest in having access to clinical trial results, including raw data, methods for analyses, detailed assessments of a study's conduct and quality, compliance with protocol, and other trial documentations. Although the regulatory authority is aware of the progress of a trial, it should not interfere with the decision made regarding the progress or termination of the trial. That is to say, the authority should not routinely participate in the DMC meeting and should not

be a voting member of DMC. The involvement of the regulatory authority with the DMC should not go beyond communication and regulatory authority's representative's attendance at open sessions so that blinding of the trials can be maintained. However, regulatory agencies should be informed of the DMC decisions as soon as possible. O'Neill (1993) indicates that this model has evolved from involvement in AIDS studies. The model has worked well for various NIH trials especially for the ATCG program of the U.S. National Institute of Allergy and Infectious Disease.

In the late 1960s, although the concept for having a formal committee being in charge of reviewing the accumulating data for safety and efficacy of medicines under investigation was well received, it was not until recently that a few trials sponsored by the pharmaceutical industry or medical community incorporated the concept of a data monitoring committee (DMC). The reasons for increasing number of industry-sponsored clinical trials employing DMC include:

1. Growing number of industry-sponsored trials with mortality or major morbidity endpoints.
2. Increasing number of trials cosponsored by the pharmaceutical industry and government that requires DMC under the policies of government funding agencies.
3. Awareness within the medical/scientific community of problems in trial conduct and analysis that might lead to inaccurate and/or bias results.

However, as mentioned above, many different models have been proposed and used for the operation of DMC under different environment, sponsors, and funding agencies with various degrees of success, failure, advantages, and drawbacks. In November 2001, the Center of Biological Evaluation and Research (CBER) of the U.S. FDA issued a draft guidance on the *Establishment and Operation of Clinical Trial Data Monitoring Committees* that discusses the roles, responsibilities, and operating procedures of DMC. Although currently the FDA regulations do not require the use of DMC in trials, this draft guidance assists sponsors of clinical trials in determining when a DMC is needed for optional study monitoring, and how such committees should operate. Various technical and ethical issues are emerging as the use of DMC and interim analysis becomes increasingly popular. Dixon and Lagakos (2000) and DeMets (2000) provide different views on whether different DMCs are allowed to share confidential data. On the other hand, Korn and Simon (1996) discussed the issues of lower-than-expected accrual or event rates in monitoring data. Reboussin et al. (2000) described an interactive Fortrain program for computation of group sequential boundaries using the Lan–DeMets alpha spending function method that does not require an equally spaced time interval and a prespecified number of interim analyses.

12.8 USE OF GENETIC INFORMATION FOR EVALUATION OF EFFICACY

As indicated in Chapter 2, one type of variation that is commonly observed in any clinical responses is the biological difference among trial subjects. Factors such as age, gender, education or social-economic status, smoking habit, weight, sexual orientation, and the underlying disease characteristics at baseline may contribute to the variation among subjects. To reduce the biological variation and to improve the accuracy and precision and generalizability of the trial results, the inclusion/exclusion criteria are usually clearly specified in the study protocol

to keep the characteristics of subjects as homogeneous as possible in the targeted patient population. In addition, other techniques such as blinding, randomized allocation of patients to treatments, and standardized procedures for evaluation of clinical outcomes are implemented to eliminate possible bias and to reduce different sources of variation. However, despite these efforts and stringent scientific and statistical requirements by various guidances such as ICH E6 guideline for *Good Clinical Practice* or ICH E9 guideline for *Statistical Principles for Clinical Trials*, considerable variation of clinical responses or outcomes still occurs for most of the clinical trials. One of the fundamental reasons may be due to the genetic differences among trial participants. For example, O'Brien and Dean (1997) reported that genes were found to protect against HIV infection. In addition, Winkler et al. (1998) observed a correlation between the variability in survival of the HIV-infected subjects with their genotypes. Weinshilboum (2003) and Evans and McLeod (2003) provided an excellent review on inheritance and drug response and on pharmacogenomics, respectively. How to use the genetic information for assessment of new treatments in clinical trials has emerged as a new challenge to all clinical scientists/researchers.

As the Human Genomic Project has come to its conclusion, the mapping and sequencing of the human genome has created a tremendously rich amount of genetic information that provides clinical researchers with a golden opportunity to gain insight and understanding of genetic mechanisms for causes of variation observed in clinical responses among subjects. The new addition of genetic information will not only revolutionize the drug development for the pharmaceutical industry, but also it will utilize the science of pharmacogenomics to individualize treatment selection and to predict efficacy and safety outcomes of selected treatments for the individuals. However, the impact of genetic information on the clinical outcomes for all new treatments for different diseases must be rigorously and scientifically studied in clinical trials before their real clinical practices. Before the genomic era, the data generated from clinical trials are basically the clinical outcomes of the phenotypic data such as reduction in sitting systolic blood pressure, fasting plasma glucose level, increase in high-density lipoprotein (HDL) cholesterol level, or prolongation of survival. However, these clinical outcomes are in fact derived from complicated interactions between genetic factors and environmental factors—treatment. Now, in the post-genomic era, the availability of genetic information enables us to link the phenotypic data with the genotypic data to investigate the mechanisms of the interaction of genes and treatments on the clinical outcomes. In this section, we will introduce three examples from real clinical trials to illustrate the use of genetic information on selection of patients and analysis of results. We will also address the impact of genetic information on sample size and duration of trial.

Example 12.8.1 Treatment of Chronic Myelogenous Leukemia

It was estimated that chronic myeloid leukemia constitutes about 20% of newly diagnosed leukemia (CML) in adults. It is a clonal disorder in which cells of myeloid lineage undergo massive clonal expansion. The disease is characterized by three distinct phases: a chronic phase of duration between three to six years, an accelerated phase, and then a blast crisis, during which the leukemia loses its differentiation ability. The only curable treatment of CML is allogeneic stem-cell transplantation. However, it is associated with considerable mortality and morbidity. In addition, only 30% of CML patients have suitably matched donors. The first line of drug therapy includes interferon-alfa that can induce a complete cytogenetic response only in about 5–20% of the CML patients with serious adverse events. The cytogenetic responses are rare for the second line agents, including hydroxyurea or busulfan after the failure of interferon-alfa.

The reason for 90% of CML patients is due to a reciprocal translocation between the long arms of chromosomes 9 and 22, which forms the so-called *Philadelphia (Ph+) chromosome*. The direct consequence of this genetic abnormality from this reciprocal translocation of regions of the *BCR* and *ABL* genes is the formation of a *BCR-ABL* fusion gene. The product of the *BCR-ABL* fusion gene is the generation of the fusion *BCR-ABL* protein, which is a constitutively activated tyrosine kinase with an important role in regulation of cell growth, and it can be found in almost all CML patients. Both *in vitro* studies and animal models have established that *BCR-ABL* tyrosine kinase alone is sufficient to induce CML. In addition, mutational analysis has shown that the tyrosine kinase activity is required for its oncogenic activity. Because of this mechanism, an inhibitor of the *BCR-ABL* tyrosine kinase could be an effective treatment for CML.

Imatinib mesylate is one of the selective and competitive inhibitors of the *BCR-ABL* protein tyrosine kinase. Several phase I and phase II trials have shown that imatinib mesylate at 400 mg per day after failure of previous treatment of interferon-alfa can induce major cytogenetic responses in about 60% of patients with confirmed late chronic phase CML and complete hematologic responses in about 95% of the CML patients (Druker et al., 2001; Kantarjian et al., 2002). Because imatinib mesylate was the first drug successfully using the concept of molecular targeting in treating cancer patients, under the *accelerated approval* regulations for severe or life-threatening illnesses, in less than three months, on May 10, 2001, the U.S. FDA approved imatinib mesylate for oral treatment of patients with CML based on the results of surrogate cytogenetic and hematologic endpoints from three separate single-arm studies in about 1,000 patients. Imatinib mesylate is also a selective inhibitor of the transmembrane receptor *KIT*, which has tyrosine activity and is the product of the *KIT* proto-oncogene. It has been reported that *KIT* activation occurs in all cases of gastrointestinal stromal tumors (GIST) (Rubin et al., 2001). Demetri et al. (2002) reported the results of a phase II study in which imatinib mesylate 400 mg or 600 mg daily cannot only induce any complete response, but also it can provide a 54% partial response rate in patients with advanced gastrointestinal stromal tumors. On February 1, 2002, the U.S. FDA also approved imatinib mesylate for oral treatment of patients with GIST. However, the real clinical benefit contributed by imatinib mesylate for the patients with either CML or GIST such as improvement on survival remains to be confirmed.

Although the genetic mechanism for CML and its molecular target are quite clear, not all patients can achieve cytogenetic or hematologic responses. When they do, the extent of responses and time required to reach responses also vary among patients. Therefore, considerable variation in responses to treatment of imatinib mesylate still exists among patients with CML. In other words, *Philadelphia (Ph+) chromosome*, *BCR-ABL* fusion gene, and its product, *BCR-ABL* tyrosine kinase, are sufficient to induce CML. There might be other causes for CML. In addition, variation of the *BCR-ABL* fusion gene and its interaction with other known or unknown genes may cause the variation in responses of CML patients to the treatment of oral imatinib mesylate.

Example 12.8.2 Treatment of Metastatic Breast Cancer

Approximately 1.6 million women have breast cancer in the United States, with 180,000 new cases each year. In addition, it is estimated that each year more than 40,000 women die of metastatic breast cancer in the United States despite recent advances in diagnosis and treatment of breast cancers (Hortobagyi, 1998). Chemotherapy can induce objective responses in most of the patients with metastatic breast cancers. However, it can rarely cure the cancer. Moreover, most forms of chemotherapy are cytotoxic and can cause serious and

substantial adverse events. The human epidermal growth factor receptor (*HER2*) is a growth factor receptor gene that is amplified in about 30% of the patients with metastatic breast cancer. *HER2* encodes the *HER2* protein found on the surface of some normal cells that plays an important role in regulation of cell growth. In addition, this encoded protein is present in abnormally high levels in the cancerous cells. Studies have demonstrated that patients with breast cancers with overexpressed *HER2* have an aggressive form of the cancer with statistically significantly shorter progression-free survival and overall survival (Seshadri et al., 1993; Ravdin and Chamness, 1995). Because the *HER2* gene is a prognostic and predictive marker in breast cancer, it provides an opportunity to target an inhibitor of *HER* protein as a treatment for the patients with metastatic breast cancer.

Herceptin is a monoclonal antibody bioengineering from part of a mouse antibody whose anti-binding region was fused to the framework region of IgG to minimize immunogeneity. It was tested against breast-cancer cancer cells that overexpressed *HER2* both *in vitro* and *in vivo*. When used alone, Herceptin was found to inhibit tumor growth. In addition, it provides additive effects when used in combination with other chemoagents such as paclitaxel. Phase I studies demonstrated that Herceptin is relative safety, and because of its selectivity, most of the antibody is confined to tumor cells. Phase II trials showed that for after the failure of previous chemotherapy, Herceptin induced an objective response in a considerable number of women with the metastatic breast cancer. Therefore, several large-scaled, randomized Phase III trials were conducted to confirm the effectiveness and safety of Herceptin in patients with overexpressed *HER2* (Slamon et al., 2001).

Enrichment design was actually employed in these studies to compare Herceptin plus chemotherapy with chemotherapy alone. Unlike the clinical endpoints such as reduction of VPC in CAST or ADAS for Alzheimer's disease, as illustrated in Chapter 5, employed during the enrichment phase, immunohistochemical assay for the levels of expression of *HER2* was used to screen the patients with overexpressed *HER2*. In particular, Slamon et al. (2001) only randomized the patients who had weak-to-moderate staining of the entire tumor-cell membrane for *HER2* (score of 2+) or more than moderate staining (score of 3+) in more than 10% of tumor cells on immunohistochemical analysis. The overall objective responses rates were 50% and 32% for Herceptin plus chemotherapy and chemotherapy alone. The one-year survival rate was 78% for Herceptin plus chemotherapy and 67% for chemotherapy alone. In addition, the patients who respond best to Herceptin had the highest levels of *HER2* protein. The most important adverse event is congestive heart failure that occurred in 27% of the patients receiving Herceptin in combination with anthracyclines and cyclophosphamide (AC). Therefore, Herceptin was not approved to be used with AC by the U.S. FDA in 1998. Unlike the trials using enrichment design discussed in Chapter 5, the enrichment phase of the trial by Slamon et al. (2001) actually can be implemented into real clinical practice to identify the 20–30% of the patients with metastatic breast cancer and with overexpressed *HER2* who can potentially benefit from treatment of Herceptin. Therefore, it is extremely critical to develop accurate, reliable, and yet inexpensive devices or procedures to achieve this goal with acceptable sensitivity and specificity. Bazell (1998) described the details of ups and downs on the development of Herceptin.

Example 12.8.3 Estrogen Receptor Polymorphism

Until the conclusive evidence provided by the Women Health Initiatives (Writing Group for WHI, 2002), elevation of plasma levels of high-density lipoprotein (HDL) cholesterol by estrogen was thought to be one of the reasons for the lower incidence rate of heart

disease in postmenopausal women receiving hormonal replacement therapy (HRT). However, considerable variation of HDL levels exists in the postmenopausal women who receive exogenous estrogen provided by HRT. A significant portion of variation of HDL levels observed among postmenopausal women receiving HRT may be due to genetic factors or to interaction between genetic factors and HRT. One of the genes responsible for variation in the phenotypic measurement is the gene located at 6q24.1 that encodes estrogen receptor α (*ER- α*). Its allelic variants can modify its expression and hence may account for some variation in the HDL level.

The Estrogen Replacement and Atherosclerosis (ERA) trial (Herrington et al., 2002) is a randomized, double-blind, three-arm, placebo-controlled trial to evaluate the effects of estrogen replacement therapy (0.625 mg per day oral conjugated estrogen) with or without continuous low-dose Progestin (2.5-mg oral medroxyprogesterone acetate per day) versus placebo on progression of coronary artery atherosclerosis in a total of 309 women with angiographical verified coronary disease (Herrington et al., 2000a; Herrington et al., 2000b). One of the objectives for the ERA trial was to measure and identify the association between 10 sequence variants in *ER- α* and the response of the HDL cholesterol to HRT among postmenopausal women. For this specific objective, the phenotypic data, i.e., HDL cholesterol level including subtraction 2 (HDL2) and subtraction cholesterol level, must be measured. In addition, to obtain the genotypic data, DNA was isolated and genotyping was performed for each single nucleotide polymorphism (SNP). A linear model was used to describe the relationship among HDL levels and estrogen treatment, various genotypes, and their interaction after adjustment for the covariates such as baseline HDL levels, age, race, diabetes status, body-mass index, baseline smoking status, frequency of exercise, and alcohol intake.

As the result of resequencing by the ERA trial, a total of nine SNPs and the TA-repeat dinucleotide polymorphism were identified and located within *ER- α* . These 10 nucleotide polymorphisms were in the Hardy-Weinburg equilibrium. In addition, frequencies for the variant SNP alleles ranged from 7.3% to 47.8% (Herrington, 2002). Results from the analysis by linear models indicated that increases in HDL cholesterol levels with HRT were greatest in women who were homozygous for the less common alleles for the intron 1 polymorphisms of human estrogen receptor α . One of the four SNPs is the first intervening sequence-401 (IVS1-401) polymorphism that involves the substitution of nucleotide from cytosine to thymine. As shown in Table 12.8.1, HDL cholesterol levels remained relatively

Table 12.8.1 Mean High-Density Lipoprotein Cholesterol Levels and Baseline and Follow-up with Respect to Treatment Group and Genotypes of IVS1-401

	Genotype of IVS1-401								
	C/C			C/T			T/T		
	HRT	Placebo	Diff.	HRT	Placebo	Diff.	HRT	Placebo	Diff.
Baseline	47.4	47.6	-0.2	47.2	44.4	2.8	42.8	44.6	-1.8
Follow-up	60.4	47.9	12.5	53.1	47.7	5.4	48.8	45.2	3.6
Difference	13.0	0.3	12.7	5.9	3.3	2.6	6.0	0.6	5.4

p-value for genotype-by-treatment interaction = 0.0009.

Summarized from Herrington et al. (2002).

unchanged for the placebo group irrespective of different genotypes. However, the responses in HDL cholesterol levels of the patients receiving the hormonal replacement therapy varied considerably with respect to different genotypes. In particular, the differences in change from baseline in HDL cholesterol levels between HRT and placebo groups for genotypes C/C, C/T, and T/T were 12.7, 2.6, and 5.4 mg/dL respectively. It followed that the interaction between the first intervening sequence-401 (IVS1-401) and treatment was greatest among the four SNPs in intron 1 with a *p*-value of 0.004 by the dominant model. The above results actually represent the differences in means of the marginal distributions for change from baseline in HDL levels between the treatment groups for genotypes C/C, C/T, and T/T averaged over all other nine nucleotide polymorphisms. Therefore, evaluation of joint effects of different *ER- α* polymorphisms and possibly with closed linked genotypes remains a challenge to clinical scientists/researchers though the augmented response of HDL cholesterol to the HRT in the women with the *ER- α* . IVS1-401 C/C genotype might be worth performing genetic screening for this allelic variant if elevation of HDL cholesterol can translate into clinical benefit of reducing the risk of cardiovascular events.

Example 12.8.4 Sample Size Considerations

From Example 12.8.1, 90% of CML patients have *Philadelphia (Ph+)* chromosome. On the other hand, from Example 12.8.2, 20–30% of the patients with metastatic breast cancer have overexpressed HER2. In addition, from Herrington et al. (2002), about 19% of the postmenopausal women have the IVS1-401 C/C genotype that induces an augmented effect on the response of HDL cholesterol level to hormonal replacement therapy. Therefore, a significant portion of disease susceptibility might be due to genetic factors. For example, Winkler et al. (1998) reported a tremendous variation of survival for HIV-infected patients with different genotypes. The 18-year survival rate ranges from 30% for the patients with less protective genes to almost 100% for the genetic protective individuals. For simplicity, one can classify the genes or their genotypes into two groups. The first group of the genes will respond to the environmental intervention such as drug therapy. This group of the genotypes may be referred to as the *reactive genotypes*. The other group of the genotypes belongs to those whose impact on the response to treatment intervention is relatively minimal. This group of the genotypes may be referred to as the *inert genotypes*. For clinical trials without including an enrichment phase for screening the subjects with reacting genotypes, heterogeneity among the subjects due to the genetic variation not only reduces the trial's power, but also it increases the study duration (Fijal et al., 2000).

Suppose that the primary endpoint for a particular trial is an occurrence of a predefined event such as eradication of a certain strain of bacteria causing infection or improvement of disease defined by some objective criteria. One measure to evaluation of the effectiveness of the treatment is to compare the response rate between treatment groups. As illustrated in Chapter 9, the response rate for a treatment group is usually estimated as the proportion of the subjects with occurrence of the event divided by the total number of subjects receiving the assigned treatment. If every patient has the reactive genotypes, then the estimated response rate is unbiased. However, if some patients contain inert genotypes, they will never respond or respond differently to the treatment regardless of whether the treatment is efficacious, and hence, the observed proportion of the subjects with occurrence of the predefined event will be underestimated for the true response rate.

Let G_R and G_I be the probabilities that a subject has the reactive or inert genotypes, respectively, and P_{Rj} and P_{Ij} denote their corresponding true response rates under treatment j , $j = T, C$. We can always assume that $P_{Rj} \geq P_{Ij} \geq 0$. Therefore, the observed response rate can be written as

$$p_j = G_R P_{Rj} + G_I P_{Ij}, \quad j = T, C. \quad (12.8.1)$$

For simplicity, we consider the extreme situation in which the patients with inert genotypes will never respond to any treatment, i.e., $P_{Ij} = 0$ for all j . p_j in (12.8.1) then can be simplified as

$$p_j = G_R P_{Rj} < P_{Rj}, \quad j = T, C. \quad (12.8.2)$$

Then the observed response rate is a biased estimator for the true response rate for the target population of the patients with reactive genotypes. If the proportion of the patients with inert and reactive genotypes is the same for both test and control groups, then the difference in observed response rates between treatment groups also underestimates the difference in the true response rates because

$$p_T - p_C = G_R(P_{RT} - P_{RC}) < P_{RT} - P_{RC}. \quad (12.8.3)$$

Underestimation of the difference in true response rates due to inclusion of patients with inert genotypes will increase the required sample size and the study duration. Screening the patients with reactive genotypes, therefore, will lead to a smaller sample size and presumably a shorter trial period. However, as shown in Examples 12.8.2 and 12.8.3, the proportion of the patients with reactive genotypes sometimes is rather smaller, usually under 30%. More subjects will be screened to obtain the required sample size of the patients with reactive genotypes. In addition, the cost for screening and the length of accrual period increase as the proportion of the subjects with reactive genes decreases. In concept, genotypic screening is very important to reach a homogenous target population for a clinical trial with a smaller sample size and a short duration. The results from such studies can be also transformed into real clinical practice for the patients who can benefit from the treatment. However, implementation of genotypic screening for a clinical trial must consider a lot of real issues, such as cost, accrual, ethnics, and privacy of the subjects. For more details on impact of genotypes on sample size and study duration, see Fijal et al. (2000). Lavori et al. (2002) reported the principles, organization, and operation of a DNA bank for clinical trials.

12.9 SAMPLE SIZE RE-ESTIMATION

In clinical trials the sample size is determined by a clinically meaningful difference and information on the variability of the primary endpoint. Since the natural history of the disease is usually not known or the test drug under investigation is a new class of drug, the estimate of variability for the primary endpoint for sample size determination may not be adequate. As a result the planned sample size may need to be adjusted in the middle of the trial if the observed variance of the accumulated responses on the primary endpoint is very

different from that used at the planning stage. Procedures have been proposed for adjusting the sample size (or re-estimation) during the course of the trial without unblinding and altering the significance level (Gould, 1992, 1995a; Gould and Shih, 1992).

For example, let us consider a randomized trial with two parallel groups comparing a test drug and a placebo. Suppose that the distribution of the response of the primary endpoint is (or is approximately) normal. Then the total sample size for a two-sided alternative hypothesis can be determined using the following formula as given in Example 11.3.2:

$$N = \frac{4\sigma^2[Z(\alpha/2) + Z(\beta)]^2}{\Delta^2}. \quad (12.9.1)$$

In general, σ^2 , the within-group variance, is unknown and must be estimated based on previous studies. Let σ^{*2} be the within-group variance specified for sample size determination at the planning stage of the trial. At the initiation of the trial, we expect the observed variability to be smaller than σ^{*2} so that the trial will have sufficient power to detect the designated clinical meaningful difference. However, if the variance turns out to be much larger than σ^{*2} , we will need to re-estimate the sample size without breaking the randomization codes. If the true within-group variance is in fact σ'^2 , then the sample size to be adjusted to achieve $(1-\beta)100\%$ power at the α level of significance for a two-tailed alternative is given by

$$N' = N \frac{\sigma'^2}{\sigma^{*2}}, \quad (12.9.2)$$

where N is the planned sample size calculated from σ^{*2} .

However, σ'^2 in (12.9.2) is unknown and must be estimated from the accumulated data available from a total n of N patients. One simple approach to estimate σ'^2 is based on the sample variance calculated from the n responses which is given by

$$(n-1)s^2 = \sum \sum (Y_{ij} - \bar{Y}_.)^2, \quad (12.9.3)$$

where Y_{ij} is the j th observation in group i and $\bar{Y}_.$ is the overall sample mean, $j = 1, \dots, n_i$, $i = T, P$, and $n = n_T + n_P$. If n is large enough for the mean difference between groups to provide a reasonable approximation to Δ , then it follows that σ'^2 can be estimated by (Gould, 1995a)

$$\sigma'^2 = \frac{n-1}{n-2} \left(s^2 - \frac{\Delta^2}{4} \right). \quad (12.9.4)$$

Note that the estimation of within-group variance σ'^2 does not require the knowledge of treatment assignment, and hence the blindness of the trial is maintained. However, this approach does depend on the mean difference, which is not calculated and is unknown. The other procedure for estimating σ'^2 without a value for Δ is the EM algorithm (Gould and Shih, 1992; Gould, 1995a). Suppose that n observations, Y_i , $i = 1, \dots, n$, on a primary endpoint have been obtained from n patients. The treatment assignments for these patients are unknown, and Y_i are the observations of patient i that can be randomly allocated to either of the two groups. Gould and Shih (1992) and Gould (1995a) assume that the treatment assignments are *missing at random*. Define π_i as the treatment indicator

$$\pi_i = \begin{cases} 1 & \text{if the treatment is the test drug} \\ 0 & \text{if the treatment is placebo.} \end{cases} \quad (12.9.5)$$

The E step is, after substitution of the current estimates of μ_T , μ_P , and σ , to obtain the provisional values of the expectation of π_i (i.e., the conditional probability that patient i is assigned to the test drug given Y_i), which is given by

$$P\{\pi_i = 1|Y_i\} = \frac{1}{1 + \exp[(\mu_T - \mu_P)(\mu_T + \mu_P - 2Y_i)/2\sigma^2]}, \quad (12.9.6)$$

where μ_T and μ_P are the population mean of the test drug and the placebo, respectively. The M step involves the maximum likelihood estimates of μ_T , μ_P , and σ after updating π_i by their provisional values obtained from (12.9.6) in the log-likelihood function of the interim observations, which is given by

$$1 = n \log \sigma + \frac{\sum [\pi_i(Y_i - \mu_T)^2 + (1 - \pi_i)(Y_i - \mu_p)^2]}{2\sigma^2}. \quad (12.9.7)$$

The E and M steps are iterated until the values converge. Gould and Shih (1992) and Gould (1995a) indicate that this procedure can estimate within-group variance quite satisfactorily, but fail to provide a reliable estimate of $\mu_T - \mu_P$. As a result the sample size can be adjusted without knowledge of treatment allocation. For sample size re-estimation with respect to binary clinical endpoints, see Gould (1992, 1995a).

From the above procedures, the sample size adjustment can be performed during the study without unblinding the treatment allocations and knowledge of any treatment differences. Therefore it is not necessary to adjust the significance level. However, the magnitude of adjustment of the planned sample size for the phase III pivotal trials may be relatively significant. This is a consequence of the variability between observations from phase II and from the current phase III trials. One has to find the possible causes for any within-group variances such as due to different patient populations between phase II and phase III trials. Although the technique does not require adjustment of the p -value and does maintain blinding, sample size re-estimation is rather seldom performed because of concerns for the integrity of a trial (Williams et al., 1993). For a review of the methods for sample size re-estimation, see Shih (2001b).

12.10 DISCUSSION

The statistical analysis of efficacy data from clinical trials involves many complicated issues. The problems of missing values and dropouts add yet another dimension to this complex analysis. There are two types of missing values (Diggle, Liang, and Zeger, 1994). The first concerns missing values when patients withdraw from a trial at any time before its planned completion. As a result the data scheduled to be collected beyond the patient's dropout time are missing. We refer to the missing values due to reasons other than dropouts as intermittent missing values.

There are many causes of dropouts and missing values. Dropouts can occur because of the duration of study, the nature of the disease, the efficacy and adverse effects of the drug under study, intercurrent illness, accidents, patient refusal or moving, and any number administrative reasons. Some of these causes are treatment-related and some are not. Based on these causes, the mechanism of missing values can generally be grouped into three types (Little and Rubin, 1987). If the causes of missing values are independent of the observed responses and of the responses which would have been available had they not been missing, then the missing values are said to be *completely random*. On the other hand,

if the causes for missing values depend on the observed responses but are independent of the scheduled but unobserved responses, then the missing values are said to be *random*. The missing values are said to be *informative* if the causes for missing values depend on the scheduled but unobserved measurements.

If the missing mechanism is either completely random or random, then the statistical inference derived from the likelihood approaches based on patients who complete the study is still valid. However, this inference is less efficient under the completely random or random missing mechanism (Diggle, Liang, and Zeger, 1994). If the missing values are informative, then the inference based on the completers would be biased. As a result the FDA and ICH guidelines on clinical reports both indicate that despite the difficulty, the possible effects of dropouts and missing values on magnitude and direction of bias must be explored as fully as possible. That is to say, before any analyses of efficacy endpoints, at various intervals during the study the frequency, reasons, and time to dropouts and missing values should also be compared between treatment groups. Their impact on the trial's efficacy has to be fully explored and understood, and only after the missing mechanism has been identified can appropriate statistical methods be employed for the analysis (Diggle, 1989; Ridout, 1991). Although some procedures have been proposed (Diggle and Kenward, 1994), there exists no satisfactory well-developed methodology to account for missing values or intermittent missing values. Therefore the conservative strategy seems to be to continue to collect the data on the primary endpoints after a patient withdraws from the study and then to analyze the efficacy endpoints based on the intention-to-treat principle with all available data from all randomized patients.

13

SAFETY ASSESSMENT

13.1 INTRODUCTION

As was indicated in Chapter 1, the safety of marketed drugs did not become a public concern until the Elixir Sulfanilamide disaster in late 1930s which led to the Federal Food, Drug and Cosmetic Act (FD&C Act). The FD&C Act requires the pharmaceutical companies to submit all reports of investigations on the safety of new drugs. This requirement was subsequently strengthened by the Kefauver-Harris Drug Amendments to the FD&C Act. These federal regulations have influenced the quantity and quality of safety information on drugs on the market. Consequently, throughout the development and marketing of a new drug, there are involved stages of government safety assessments. O'Neill (1988) classifies these stages as pre-marketing, post-marketing, and drug labeling. O'Neill indicates that the assessment of the safety of a new drug begins in the pre-approval stage with pre-clinical animal studies and with early phase I studies that examine the absorption, excretion, dose ranging, tolerance, and other pharmacokinetic performance of the drug in humans. This continues with phase II and phase III of clinical development. During the post-approval marketing, much broader patient populations become involved. The safety information may be obtained from voluntary reports, monitoring system, uncontrolled patient follow-up, and formal epidemiological studies.

For the pre-marketing safety assessment, the FDA requires that a summarization and analysis of safety information of a new drug be included in the NDA submission. In particular, Section 314 of CFR [314.50 (d)(5)(ii)] indicates that an integrated summary of all available information on the safety of the drug product be submitted, including pertinent animal data, demonstrated or potential adverse effects of the drug, clinically significant drug-to-drug interactions, and other safety considerations such as data from the epidemiological

studies of a related drug. The applicant has to date periodically its pending application with new safety information learned about the drug that may reasonably affect the statement of contraindication, warnings, precautions, and adverse reactions in the draft labeling. After the drug is approved, Section 21 CFR 314.80(b) also requires that the sponsor promptly review all adverse drug experience information obtained or otherwise received by the sponsor from any source, foreign or domestic, including information derived from commercial marketing experience, post-marketing clinical investigations, post-marketing epidemiological/surveillance studies, reports in scientific literature, and unpublished papers. Section 21 CFR 314.80(c)(1)(ii) also requires that the sponsor periodically review the frequency of adverse drug experience reports which are both serious and expected (in the labeling) and report any significant increase in frequency that might suggest a drug-related incidence higher than previously observed or expected. The safety update report should be submitted (1) four months after the initial submission, (2) following receipt of an approvable letter, and (3) at the times requested by the FDA. For drug labeling, Section 21 CFR 201.57(e) on warnings states when the label should describe serious adverse reactions and potential safety hazards. In addition, Section 21 CFR 201.57(g)(2) requires that the frequency of the serious adverse reactions be expressed under the section of *Adverse Reaction* of the labeling, and if known, the approximate mortality and morbidity rates for patients sustaining the reaction.

In general, the assessment of drug safety in clinical trials has not received the same level of attention as the assessment of efficacy. For example, in most clinical trials, sample size is determined to achieve a desired power for detection of a clinically meaningful difference in the primary efficacy variable rather than safety parameters. In addition, unlike the assessment of efficacy, the hypotheses for safety assessment are usually much less well-defined. As a result statistical methods for the assessment of safety are limited. In practice, descriptive statistics for safety data obtained from clinical trials, both controlled and uncontrolled, are often used to (1) summarize rates of occurrence of adverse events in exposed groups and (2) examine any patterns or trends for subgroups of patients experiencing differential rates of adverse events. In practice, safety data obtained in clinical trials are often summarized in terms of rates or relative risks of certain events. Therefore it is important to develop a sound statistical methodology that provides an accurate and reliable assessment of drug safety. In addition to safety data of rates and relative risks of certain events, most clinical trials also contain a large battery of routine laboratory measures as part of the safety evaluation of the drug under investigation. Since laboratory data are obtained from analytical methods, they are often the least biased and the most precise data collected in clinical trials. Although laboratory data can provide information on systemic toxicity, these data are often underutilized for evaluation of drug safety.

In this chapter, our goal is to describe approaches for assessing the safety of a drug product in terms of rates or relative risks of adverse events and laboratory data during its clinical development and marketing stages. In the next section the toxicity (or risk) of the exposure to a drug under study is briefly described including the incidence rate of an event and laboratory tests. In Section 13.3 we provide definitions for adverse drug reaction, adverse event, and serious adverse events adopted by the FDA and ICH. Also included in the section are the coding, filing, and reporting of observed adverse events according to some acceptable dictionaries such as COSTART, MedDRA, or IMT. Statistical methods for the assessment of safety in terms of adverse events are reviewed in Section 13.4. Analyses of laboratory data for the assessment of safety are discussed in Section 13.5. Some remarks and discussions are given in the last section.

13.2 EXTENT OF EXPOSURE

As indicated in *Federal Register* (vol. 61, no. 138, 1996) and ICH E3 guideline entitled Structure and Content of Clinical Study Reports (1996), the evaluation of safety-related data can be considered at three levels: (1) the extent of exposure, (2) the more common adverse events and laboratory test changes, and (3) serious adverse events and other significant adverse events. In this section we will discuss the extent of exposure of a test drug or an investigational drug product. The extent of exposure can be used to determine the degree to which safety can be assessed from the study. The extent of exposure to a test drug or an investigational product is usually characterized according to the number of patients exposed, the duration of exposure, and the dose to which patients were exposed. Duration of exposure to any dose is usually expressed as a median or mean. In practice, it is helpful to describe the number of patients exposed for specified periods of time. In addition it is suggested that the number of patients exposed to the test drug for various durations be broken down by age, sex, racial subgroups, and any other pertinent subgroups (e.g., disease, disease severity, or concurrent illness) in order to get a good profile of the exposure effect on different populations. Another measure of the extent of exposure is the number of patients exposed to specified daily dose levels. The daily dose levels used could be the maximum dose for each patient, the dose with longest exposure for each patient, or the mean daily dose. Similarly it is suggested that the number of patients exposed to various doses be broken down further by age, sex, racial, and any other pertinent subgroups to examine the profile of the extent of exposure for dose. In some cases a cumulative dose may be pertinent. In practice, it is often useful to provide combined dose-duration information such as the numbers exposed for a given duration to the most common dose, the highest dose, or the maximum recommended dose.

Risk of Exposure

In clinical trials the toxicity or risk of exposure to a drug can generally described as a function of the exposure to the drug:

$$\text{Toxicity} = f(\text{exposure}),$$

where the exposure to the drug depends on the dose and time of exposure. If we assume a constant dose, then the toxicity is a function of the time of exposure. The toxicity or risk of exposure is usually measured by some parameters such as occurrence, number, duration, and time pattern of an event. The event may be an absorbing (irreversible) event (e.g., death), a recurring even with negligible duration (e.g., seizure), or a recurring event with a duration (e.g., migraine headache episodes). Therefore, all of the parameters such as occurrence, number of events, and duration are not meaningful for all events. Meaningful parameters are rather the probability of occurrence, the number of events, and the percentage of time affected for an absorbing event, a recurring event, and a recurring event with a duration, respectively.

For quantification of the exposure risk of patients to a drug, the most commonly used measure is a crude incidence rate, which is defined as

$$CR = \frac{\text{Number of patients with the event}}{\text{Number of patients at risk}},$$

where the *patients at risk* are all patients enrolled and treated. Under the assumption that each patient has the same probability of experiencing the event, *CR* is a binomial estimate. In practice, the binomial holds only if each patient undergoes one exposure unit. Tremmel (1996) points out that even by this assumption, the incidence rate may be misleading in long-term clinical trials because (1) it does not take exposure time into account and (2) every patient experiencing the event must disregard parameters of severity such as duration and the number of events. Further, in cases where there are early terminations, the full initial patient sample used as the denominator will lead to a downward bias. Thus Tremmel (1996) suggests that *CR* be used only when each patient receives one exposure unit such as in phase I trials. For short-term trials the use of *CR* is also justifiable if all patients experience about the same amount of exposure.

As an alternative to the *CR* estimate, Tremmel (1996) suggests the following two basic techniques which can be used to quantify the exposure risk of patients to a drug by controlling for the exposure to the drug. The first is to divide the number of events by total exposure by implicitly assuming a constant hazard. In other words, we calculate the number of events per time units such as the number of deaths per patient year. The other is to consider stratification by exposure time by forming time intervals. In other words, we consider estimates of the number of events per time interval assuming a constant hazard function.

Absorbing Events

For describing the exposure risk for absorbing events, Tremmel (1996) suggests using the number of events per patient-time unit, which is defined as

$$h = \frac{\text{Number of positive patient-time units}}{\text{Number of all patient-time units}}.$$

The basic assumption is that one unit of patient-time is equivalent to and independent of another unit of patient-time. In other words, each patient-time unit is a binomial observation with the probability of success $P = h$. If the absorbing event is death, Rothman (1986) estimates the risk of exposure by means of the number of deaths per time unit of exposure:

$$h = \frac{\text{Number of deaths}}{\text{Number of total exposure}},$$

where h estimates the constant hazard. This is a typical example of constant hazard, also known as death per patient year (DPPY). In practice, since the hazard may vary over time, we need to allow for a nonconstant hazard, such as a decreasing hazard function (O'Neill, 1988). For nonconstant hazards, h is a function of time (known as hazard function). As indicated by Salsburg (1993), hazard functions provide useful information that helps us not only to assess drug causality but also to identify periods of increased risks. In addition the corresponding survival rates

$$S(t) = \prod [1 - h(t)]$$

provide probability statements of occurrence similar to the crude incidence rate which are unbiased because they are based on the corrected denominators (O'Neill, 1988a). Besides, they are adjusted for exposure by being functions of exposure time.

As was pointed by Tremmel (1996), event per patient-time unit such as death per patient year is a meaningful parameter for the assessment of exposure risk if the event can only occur once.

Recurring Events with Negligible Duration

Recurring events with negligible duration are referred to as events with relevant recurrences such as reinfections, seizures, or sensitivity reactions. Similar to the absorbing events with constant hazard, one may consider the following for recurring events:

$$h = \frac{\text{Number of positive patient-time units}}{\text{Number of all patient-time units at risk}}.$$

Thus h is the risk of getting the event for the first time or again. h provides a probability estimate under the assumption that each patient-time unit at risk is a binomial observation with the probability of success $P = h$.

For nonconstant hazards, since the recurring events may be dependent on individual subjects, the quantification of the exposure risk is more complicated. Basically there are two approaches with so-called subject-induced dependencies. For models that ignore subject-induced dependencies, the hazard function $h(t)$ can be estimated within different time intervals by the number of events in that interval divided by the total patient-exposure in that interval. The expected event counts for an individual at risk can be derived by integrating the hazard function for the risk interval. Confidence intervals can also be constructed (Andersen et al., 1993). Based on the same idea, Andersen and Gill (1982) considered Cox's proportional hazards model with the hazard being a function of time and a subject-related risk score to allow for recurring events. To account for the number of preceding events, Andersen and Gill (1982) further considered including the number of preceding events as a time-dependent covariate in the model. As a result the estimation of the expected event counts becomes more difficult and depends on the stochastic structure of the future development of the covariates (Andersen et al., 1993).

To account for subject-induced dependencies, we can consider either a normal model for count data proposed by Hoover (1996) or a random-effects model (or frailty model). For the normal model for count data, the idea is to form time intervals with complete exposures by deleting data that are censored in the interval and then, for each subject and interval, determine the event count. We next estimate the mean counts, variances, and covariances of counts between time intervals, assuming that the matrix of counts is from a multivariate normal distribution. We apply the central limit theorem to justify the estimates, which we finally use to derive point estimates and confidence intervals for event counts over time. The purpose of the random effects model is to examine whether there is a subject effect over a potentially relevant variable such as exposure time. The subject effect is defined as the observed count over time for each subject.

Various statistical procedures have been proposed in the literature for analysis of recurrent events or multiple failure time data. For example, Wei and Glidden (1997) provided an overview on statistical methods for multiple failure time data in clinical trials, which were developed before 1995. These methods include the Anderson-Gill (AG) model and the methods proposed by Prentice et al. (PWP, 1981) and Wei et al. (WLW, 1989). The AG model is a generalization of the Cox proportional hazard model, which is in a multiplicative manner that relates the intensity function of recurrence to the covariates. The method

by Prentice et al. (1981) is an alternative for analysis of the recurrent events. The PWP model handles the dependence between event times by the stratification of the prior number of failures. The method proposed by Wei et al. (1989) is to obtain the maximum partial maximum likelihood estimates using the usual proportional hazards model. In addition, Pepe and Cai (1993), Lawless et al. (1997), and Lin et al. (2000) proposed a robust semi-parametric procedure for inferences about the rate and mean functions of the recurrent events. Recent development for analysis of recurrent events can be found in Wang and Chang (1999), Chang and Wang (1999), Wang and Chen (2000), Kelly and Lim (2000), Chang (2000), Lin and Ying (2001), Wang et al. (2001), and Mahe and Chevret (2001). However, little information on the power and sample size determination for the recurrent events is available (see, e.g., Hughes, 1997).

Recurring Events with Nonnegligible Duration

To quantify the risk of exposure of recurring events with nonnegligible duration, the most commonly employed approach is to calculate either the prevalence rate or the incidence rate at a given time. The prevalence rate is the risk of *having* an event at a given time, which is defined as

$$h_P = \frac{\text{Number of patient-time units affected}}{\text{Number of patient-time units exposed}}.$$

As can be seen from the above definition, the prevalence rate includes patients in both the numerator and denominator throughout the duration of an event. As a result, the prevalence rate is the proportion of a population that is affected by disease at a given point in time. Continued suffering increases the number of patient-time units affected. On the other hand, the incidence rate is the risk of *getting* an event at a given time which is defined as

$$h_I = \frac{\text{Number of incidences}}{\text{Number of patient-time units at risk}}.$$

The differences between the prevalence rate and the incidence rate are that (1) the continued suffering does not count in numerator and that (2) patient-time units of continued suffering are not at risk of getting the event and hence are removed from the denominator. Let us consider the example listed in Table 13.2.1 which was given in Tremmel (1996). In the table the total patient-time units is given by $10(4) = 40$ and the patient-time units affected is 10 (the number of crosses). Therefore, the prevalence rate is given by

$$h_P = \frac{10}{40} = 0.25.$$

On the other hand, the patient-time at risk is $40 - 6 = 34$ and the incidences is 4. Thus the incidence rate is

$$h_I = \frac{10 - 6}{40 - 6} = \frac{4}{34} = 0.12.$$

Note that the prevalence rate can be estimated by the incidence rate if it is multiplied by the average duration of the disease (Rothman, 1986).

Table 13.2.1 Recurring Events with Nonnegligible Duration

Patient	1	2	3	4	5	6	7	8	9	10
1	—	—	—	—	×	×	×	×	—	—
2	—	—	—	—	—	×	—	—	—	—
3	—	—	—	—	—	—	—	—	—	—
4	—	—	—	×	×	—	—	×	×	×

Source: Tremmel (1996).

Note: × indicates affected.

Laboratory Data

Another approach that is commonly employed to investigate the extent of exposure to a drug in clinical trials is to perform routine laboratory tests. In practice, many clinical studies contain a large battery of routine laboratory measurements as part of the safety evaluation of the drug under investigation. Compared to the rates or relative risks of certain events observed from clinical trials, laboratory data are often the least biased and the most precise data because they are obtained by analytical methods. The laboratory data not only provide reliable information on system toxicity but also valuable information for evaluating the efficacy and safety of the drug under study.

In clinical trials, routine laboratory tests are usually performed not only to screen patients for inclusion in trials prior to randomization but also to protect patients during trials. In many clinical trials, laboratory tests are considered as the primary efficacy variables and hence are used for assessment of drug effects. For example, laboratory tests for cholesterol and triglycerides are often used to assist the diagnosis of patients with coronary heart disease. Cholesterol measurements includes total cholesterol (TOTAL-C), high-density lipoprotein cholesterol (HDL-C), and low-density lipoprotein cholesterol (LDL-C). A high value of LDL-C, a low value of HDL-C, or a high value of triglycerides combined with a low value of HDL-C constitutes a high-risk factor of coronary heart disease. For a normal subject, its total cholesterol, HDL-C, LDL-C, and triglycerides should be within the ranges of 130–200 mg/dL, 35–65 mg/dL, <130 mg/dL, and <250 mg/dL, respectively. The relationship among TOTAL-C, HDL-C, LDL-C, and triglycerides can be expressed by the well-known Fieldewald equation as follows:

$$\text{LDL-C} = \text{TOTAL-C} - [\text{HDL-C} + 0.2(\text{triglycerides})].$$

For another example, consider a laboratory test for glucose for patients with diabetes mellitus. According to the American Diabetes Association, a subject with fasting glucose > 126 mg/dL or two-hour post cibum (2-hr PC) glucose > 200 mg/dL would be diagnosed as a patient with diabetes mellitus. If 2-hr PC glucose is between 140 and 200 mg/dL, then the subject is considered to have impaired glucose tolerance. In general, it is estimated that approximately 1–5% of the patients with impaired glucose tolerance will have diabetes mellitus.

In clinical trials any significant change in laboratory measurements could be an indication of potential toxicity or an exposure risk of the drug under study. A careful assessment of such changes is necessary. In general, different laboratory tests can be performed to meet different study objectives of clinical trials with different indications and patient populations. On average, about 20 to 35 laboratory tests are typically performed in a clinical trials. Tables 13.2.2 through 13.2.4 list some commonly performed laboratory tests for

Table 13.2.2 Laboratory Tests for Hematology

Laboratory Test	Normal Range
WBC (white blood cell count)	3.6–9.8 × 10 ³ /μL
RBC (red blood cell count)	Male: 4.2–6.2 × 10 ⁶ /μL Female: 3.7–5.5 × 10 ⁶ /μL
Hemoglobin	Male: 12.9–17.9 g/dL Female: 11.0–15.6 g/dL
Hematocrit	Male: 38–53% Female: 33–47%
WBC classification	
Band neutrophil	0–3% or 0–5%
Segmented neutrophil	45–70%
Lymphocyte	25–40% or 2.4 ± 0.8 × 10 ³ /μL
Monocyte	2–8%
Eosinophil	1–3% or 70–400/μL
Basophil	0–0.5%
MCV (mean corpuscular volume)	82–98 fL (i.e., 10 ⁻¹⁵ L)
MCH (mean corpuscular hemoglobin)	27–32 pg (i.e., 10 ⁻¹² g)
MCHC (mean corpuscular hemoglobin concentration)	31–36%
Platelet count	120–400 × 10 ³ /μL
RDW (red cell distribution width)	11.5–14.5
MPV (mean platelet volume)	9.8 ± 1.2 fL
Reticulocyte count	0.5–1.5%
ESR (erythrocyte sedimentation rate)	Male: <10 mm/hr Female: <20 mm/hr
Bleeding time	3–10 min
Clotting time	8–10 min
PT (prothrombin time)	10–13 s
APTT (activated partial thromboplastin time)	26–36 s
G-6-PD (glucose-6-phosphatase dehydrogenase)	4.10–7.90 IU/g Hb
Fibrinogen	200–400 mg/dL
FDP (fibrinogen degradation product)	<10 μg/mL
Total eosinophil count	70–400/mm ³

hematology, clinical chemistry, and urinalysis and their corresponding reference ranges (or normal ranges) for normalities (if available).

13.3 CODING OF ADVERSE EVENTS

13.3.1 Definitions of Adverse Events

For clinical trials, it should be recognized that the safety information for the treatment under study is at least as important as or more important than the efficacy. However, different terms and/or definitions are often employed in different trials for evaluation of treatments across different therapeutic areas. For example, adverse events, adverse experience, adverse drug reactions, side effects, severe adverse events, significant adverse events, or

Table 13.2.3 Laboratory Tests for Clinical Chemistry

Laboratory Test	Normal Range
Liver function tests	
ALP (alkaline phosphatase)	65–272 IU/L
AST/SGOT (serum glutamic oxaloacetic transaminase)	15–35 IU/L
ALT/SGPT (serum glutamic pyruvate transaminase)	3–30 IU/L or 8–45 IU/L
γGT (gamma glutamyl transferase)	5–40 IU/L
Bilirubin	0.3–1.0 mg/dL
LDH (lactic acid dehydrogenase)	150–400 IU/dL
Total protein	6.6–8.1 gm/dL
Albumin	3.9–5.1 gm/dL
Globulin	2.3–3.5 gm/dL
Renal function tests	
BUN (blood urea nitrogen)	5–20 mg/dL
Creatinine	0.7–1.5 mg/dL
Creatinine clearance	Male: 62–108 ml/min Female: 57–78 ml/min
Electrolytes	
Sodium (Na^+)	135–140 mmol/L
Potassium (K^+)	3.5–5.0 mmol/L
Chloride (Cl^-)	98–108 mmol/L
Calcium (Ca^{2+})	2.1–2.6 mmol/L
Phosphorus (P)	2.5–4.5 mg/dL
Magnesium (Mg^{2+})	1.9–2.5 mg/dL
Uric acid	Male: 3.5–7.9 mg/dL Female: 2.6–6.0 mg/dL
CPK (creatinine phosphokinase)	37–289 IU/L
Aldolase	1.7–4.9 units/L
Amylase	Serum: 30–200 IU/L Urine: 4–30 IU/2 h Lipase
Cholesterol	<200 units/L
Total cholesterol	130–200 mg/dL
HDL-cholesterol	35–65 mg/dL
LDL-cholesterol	<130 mg/dL
Apo A-1 (apolipoprotein A-1)	Male: 66–151 mg/dL Female: 75–170 mg/dL
Apo B (apolipoprotein B)	Male: 49–124 mg/dL Female: 26–119 mg/dL Triglycerides
Glucose	<250 mg/dL
AC glucose	70–110 mg/dL
30 PC glucose	90–160 mg/dL
1-hr PC glucose	90–160 mg/dL
2-hr PC glucose	75–125 mg/dL
3-hr PC glucose	70–110 mg/dL
HbA _{1C} (glycosylated hemoglobin)	4–7%
Serum iron	Male: 89–200 $\mu\text{g}/\text{dL}$ Female: 70–180 $\mu\text{g}/\text{dL}$

Table 13.2.3 (Continued)

Laboratory Test	Normal Range
Ferritin	Male: 27–300 ng/ml Female: 10–130 ng/ml
Acid P-tase (acid phosphatase)	Male: 4.7 IU/L Female: <3.7 IU/L
Protein electrophoresis	
Total protein	5.9–8.0 g/dL
Albumin	4.0–5.5 g/dL
Alpha-1 globulin	0.15–0.25 g/dL
Alpha-2 globulin	0.43–0.75 g/dL
Beta globulin	0.50–1.00 g/dL
Gamma globulin	0.60–1.30 g/dL
Lipoprotein electrophoresic	
Pre-beta	20 ± 6%
Beta	50 ± 5%
Alpha	36 ± 7%
Hemoglobin electrophoresis	
H _b A	97%
H _b A ₂	1.5–3.5%
H _b F	<2.0%
H _b C	0
H _b S	0
Osmolality	280–295 mOsm/kg

serious adverse events, risks, or toxicities are terminologies often used (either exchangeably or differently) for safety assessment in different trials across different therapeutic areas. In addition, even for the same terminology, different definitions may be applied in different therapeutic areas by different countries or regions. Nickas (1995) classified adverse event data into three categories, namely, (1) known adverse drug reaction; (2) adverse events where a causal relationship is uncertain, a possible adverse drug reaction; and (3) adverse events that are considered unrelated to study drug. Northington (1996) provided a general definition, which considers an adverse event *any negative event* that a patient/subject experiences during the course of a clinical trial. The term negative event is broad and vague and allows too much flexibility for an accurate and reliable interpretation. As a result, although this definition may increase the number of events reported, it may also increase the chance of capturing all potentially important events. More specifically, Northington (1996) indicated that an adverse event might be defined as any unfavorable change in the structure (signs), function (symptom), or chemistry (laboratory data) of the body temporally associated with participation in the clinical trials, irrespective of the believed relationship to the study drug. This specific definition would also include intercurrent illness or injuries, clinical significant results from laboratory tests or other medical procedures, and clinically significant findings uncovered during a physical examination.

On the other hand, in the 1988 FDA guideline entitled, *Guideline for the Format and Content of the Clinical and Statistical Section of an NDA*, the FDA states that an adverse event tabulation of particular interest that should be produced for all studies would include all new adverse events (i.e., those not seen at the baseline or that worsened during treatment).

Table 13.2.4 Laboratory Tests for Urinalysis

Laboratory Test	Normal Range
Dipstick tests	
pH	4.6–8.0
Protein	<8 mg/dL
Glucose	
Ketone	
Occult blood	
Urobilinogen	0.1–1.0 EU/dL
Leukocyte esterase	
Nitrite	<10 ⁵ colony/ml
Sediment	
RBC	<5/HPF
WBC	<5/HPF
Epithelial cells	0
Casts	0/LPF
Crystal	
Microorganisms	
Parasites	
Spermatozoa	
Specific gravity	1.016–1.022
Gram stain	
Bence-Jones protein	
Paragquat test	
Porphobilinogen	
Myoglobin	
Pregnancy test	
Fractional urinalysis	

Adverse events of any kind are known as treatment-emergent adverse events (TEAE). This definition, however, is also vague and subject to various interpretations. Northington (1996) provided a more explicit definition of TEAE as follows. An adverse event that occurs during the active phase of the study will be considered as a TEAE if (1) it was not present at the time the active phase of the study began and it is not a chronic condition that is a part of the patient's medical history, or (2) it was present at the start of the active phase of the study or as part of the patient's medical history but the severity or frequency increased during therapy.

Basically, an adverse event can be gathered either (1) during a clinical trial or (2) spontaneously from reports on drugs already on the market. The purpose of collecting an adverse event is to enable a complete and accurate summarization of adverse events that can be expected in the target patient population. The information can also be used to guide the practicing physician in the use of the drug for good medical practice. In addition, the events reported by a patient are helpful in determining whether he/she is likely to be related to the drug. The goal of obtaining reports of spontaneous events is to detect marked changes in frequency and seriousness of events from what were observed from the trials conducted during the clinical development. Therefore, it is extremely important to have a unique definition and terminology, as well as procedures to ensure uniform Good Clinical Practice (GCP) standards for collection, reporting, and analysis of safety information such as adverse events. Scherer and Wiltse (1996) emphasized that the definition of an adverse

event should encompass the concepts of (1) any undesirable experience (2) that occurs in clinical trial participant (3) whether or not it is considered related to the study drug, (4) even if the patient never receives the study drug (intention-to-treat). For regulatory authorities, however, the ICH has harmonized various terminologies and their definitions of the safety information in ICH E2A guideline entitled, *Clinical Safety Data Management: Definitions and Standards for Expedited Reporting* (ICH, 1995), and ICH E6 guideline entitled, *Good Clinical Practice: Consolidated Guidance* (ICH, 1996). The following are the definitions of adverse events (AE), adverse drug reactions (ADR), serious adverse events (SAE), and unexpected adverse drug reaction (UADR) given by the ICH.

Adverse Events (AE) An adverse event is any untoward medical occurrence in a patient or clinical investigational subject administrated a pharmaceutical product and that does not necessarily have a causal relationship with this treatment. An AE can therefore be any unfavorable and unintended sign (including an abnormal laboratory finding), symptom, or disease temporally associated with the use of a medicinal (investigational) product, whether or not it is related to the medicinal (investigational) product.

Adverse Drug Reactions (ADR) In the preapproval clinical experiences with a new medicinal product or its new usages, particularly as the therapeutic dose(s) may not be established, all noxious and unintended responses to a medical product related to any dose should be considered adverse drug reactions. The phrase “responses to a medicinal product” means that a causal relationship between a medicinal product and an adverse event is at least a reasonable possibility; i.e., the relationship cannot be ruled out. Regarding marketed medicinal products: A response to a drug that is noxious and unintended and that occurs at a dose normally used in humans for prophylaxis, diagnosis, or therapy of diseases or for modification of physiological function.

Serious Adverse Events (SAE) or Serious Adverse Drug Reaction (SADR) A serious adverse event or reaction is any untoward medical occurrence that any dose:

- Results in death
- Is life-threatening
- Requires inpatient hospitalization or prolongation of existing hospitalization
- Results in persistent or significant disability/incapacity
- Is a congenital anomaly/birth defect

Unexpected Adverse Drug Reaction (UADR) In an adverse reaction, the nature or severity is not consistent with the applicable product information (e.g., Investigator’s Brochure for an unapproved investigational medical product or package insert/summary of product characteristics for an approval product).

In the past, the term *side effect* has been used in different ways to describe not only negative (unfavorable) effects but also positive (favorable) effects. Therefore, the ICH E2A guideline recommends that this term no longer be used and particularly should not be regarded as synonymous with adverse event or adverse reaction. In addition, one should distinguish the difference between severe and serious adverse events. As indicated in the ICH E2A guideline, the term *severe* is used to describe the intensity of a specific event. However, this event may be of a minor medical significance, e.g., severe headache. On the

other hand, the definition of SAE is based on patient/event outcome or action criteria usually associated with events that pose a threat to a patient's life or functioning. In addition, the ICH E3 guideline entitled, *Structure and Contents of Clinical Study Reports*, also requires reporting other significant adverse events that are defined as marked hematological and other laboratory abnormalities (other than those meeting the definition of serious) and any events that led to an intervention, including withdrawal of test drug/investigational product treatment, dose reduction, significant additional concomitant therapy, other than those reported as serious adverse events. Although the ICH's various guidelines or guidances provide uniform definitions and terminologies for adverse events or adverse drug reactions for the regulatory setting, they do not provide a precise definition for grading the intensity associated with an adverse event nor the degree of causality with the study drug.

As mentioned in Chapter 6, two special features must be considered for evaluation of treatments for treating cancer patients. The first is that the patients in cancer clinical trials are those with malignant tumors. Unlike other diseases, most cancers are life-threatening diseases in which the disease process is usually irreversible, and in most cases, they are neither curable nor controllable. The second feature is that most of the anti-cancer drugs are cytotoxic agents that can generate severe, irreversible, life-threatening, and sometimes fatal outcomes, such as immuno-suppression, hepatic, renal, or cardiac toxicity. As a result, the Cancer Therapy Evaluation Program (CTEP) of the United States National Cancer Institute (NCI) and the European Organization of Research and Treatment of Cancers (EORTC) employ similar but different definitions, terminologies, and criteria for evaluating safety of cancer treatments. In the CTEP's Common Toxicity Criteria Manual (CTC manual, 1999), an adverse event is defined as any unfavorable symptoms, sign, or disease (including and abnormal laboratory finding) temporally associated with the use of a medical treatment or procedure that may or may not be considered related to the medical treatment or procedure. The definition of toxicity given in the CTC manual is an adverse event that has an attribution (the relationship to investigational agent) of possible, probable, or definite. However, because toxicity is not clearly defined by regulatory organization, the NCI recommends that the term *toxicity* not be used. For their sponsored trials, both the NCI and EORTC employ the CTC criteria, Version 2.0 (1999), for providing descriptive terminology of adverse event reporting as well as the grading severity and causality of adverse events. Although the CTEP, CTC v2.0 uses the term *toxicity* for historical reason, it recommends that the term *adverse event* with its attribution be used whenever possible. In addition, the NCI also defines the life-threatening adverse event as any adverse event that places the patient or subject, in view of the investigator's, at immediate risk from the reaction. An unexpected adverse event is any event that is not listed in the NCI Agent Specific Expected Adverse Event List. A comparison between the ICH guidelines and the NCI/EORTC guidelines reveals that (1) similar but slightly different definitions for adverse events and unexpected adverse event are given, (2) both use the same definition for serious adverse event, (3) the NCI/EORTC guidelines do not provide any definition of adverse drug reaction, (4) the ICH guideline does not provide the definition for life-threatening adverse event, and (5) the NCI/EORTC guidelines do not give the definition for other significant adverse events.

For fatal or life-threatening, unexpected adverse events or adverse drug reactions occurring during clinical investigation, the ICH E2A guideline recommends that regulatory agencies should be notified as soon as possible but no later than 7 calendar days after first knowledge by the sponsor that a case qualifies, followed by as complete a report as possible within 8 additional calendar days. On the other hand, serious, unexpected AEs or ADRs

that are not fatal or life-threatening must be filed as soon as possible but no later than 15 calendar days after first knowledge by the sponsor that the case meets the minimum criteria for expedited reporting. The ICH E2A guideline also suggests that for regulatory purposes, initial reports should be submitted with the above-prescribed time as long as the following minimum criteria are met:

1. An identified patient.
2. A suspect medicinal product.
3. An identifiable reporting source.
4. An event or outcome that can be identified as serious and unexpected, and for which, in clinical investigation cases, there is a reasonable, suspected causal relationship.

In addition, follow-up information should be actively sought and submitted as it becomes available.

Most regulatory agencies such as the U.S. FDA, the Medical Control Agency of the United Kingdom (U.K. MCA), or the MHWL of Japan recommend that a dictionary for grouping similar events be used in reporting the adverse events. The purpose of grouping the observed adverse events is to assess the safety profile of the drug under investigation. There are many dictionaries available for this purpose. These dictionaries include the *Coding Symbols for a Thesaurus of a Adverse Reaction Term* (COSTART) developed by the U.S. FDA, the *World Health Organization Adverse Reaction Terminology* (WHOART), recommended by the World Health Organization, the *Common Toxicity Criteria* (CTC) by the U.S. NCI/EORTC for cancer trials, the *International Classification of Diseases* (ICD) adopted codes, *Japanese-Adverse Reaction Terminology* (J-ART), HARTS, a dictionary developed by the then Hoechst, and most recently, the *Medical Dictionary for Regulatory Activities* (MedDRA) Terminology developed by the ICH. Among these dictionaries, coding systems, and criteria for evaluation and reporting of adverse events, the COSTART, CTC, and MedDRA are the most commonly used dictionaries and coding systems for grouping similar events in the summarization and reporting of adverse events. What follows provides a brief description of these three dictionaries and criteria.

13.3.2 COSTART

The COSTART is the terminology developed and used by the FDA for the coding, filing, and retrieving of adverse reaction reports. It provides a method to deal with the variation in vocabulary used by those who submit adverse events to the FDA. The COSTART dictionary was derived in the 1960s from the dictionary of adverse reaction terms (DART). As of the fifth edition, the COSTART has more than 6,000 glossary terms collapsing to approximately 1,200 unique COSTART terms. Basically the COSTART is divided into seven indexes that provide different information on adverse events' coding, filing, and reporting. The use of these indexes is briefly summarized below.

Index A of the COSTART dictionary is composed of a hierarchical *Body System Classification*, *General Search Categories*, and *Special Search Categories*. Hierarchical Body System Classification is the primary category, and it contains 12 subcategories which are summarized in Table 13.3.1. For each primary category there may be a number of sub-categories associated with it. The coding symbols are chosen to reflect both the primary category and its subcategories. For example, for *Body as a Whole*, the primary category is

Table 13.3.1 Body System Classification of COSTART Dictionary

Classification	Coding Symbol
Body as a whole	BODY
Cardiovascular system	CV
Digestive system	DIG
Endocrine system	ENDO
Hemic and lymphatic system	HAL
Metabolic and nutritional disorders	MAN
Musculoskeletal system	MS
Nervous system	NER
Respiratory system	RES
Skin and appendages	SKIN
Special sense	SS
Urogenital system	UG

coded as *BODY* and its subcategories Head, Neck, and Thorax are coded as *BODY/HEAD*, *BODY/NECK*, and *BODY/THOR*, respectively. The purpose of the General Search Categories is for search and retrieval strategies. The Body System Classification sometimes serves as the basis of the search strategy. For the most part the need has been superseded by the use of the pathophysiologic classification of COSTART as presented in Index F. The Special Search Categories are useful for the assessment of possible fetal and neonatal disorders associated with drug or biologic product used *in utero* or in early life. The Special Search Categories include codes for recording whether a suspected drug was given to (1) either parent before conception, (2) the mother during gestation, (3) the mother at a particular time during delivery or while nursing, (4) the fetus directly *in utero*, (5) the infant during the first two years of life. These symbols signify that a particular reaction occurred during a particular time, fetal life, or in the new born.

Index B of the COSTART dictionary gives a comprehensive alphabetical listing of the COSTART symbols and obligatory Body System categories and subcategories, and Index C lists COSTART symbols stratified by Body System categories. Index C is useful in reviewing and/or determining the selection of a COSTART term. Index D of the COSTART dictionary is a glossary that contains nearly 6,000 reported terms with appropriate COSTART symbols. Index E of the COSTART dictionary reflects the FDA's attempt for a pathophysiologic classification of COSTART terminology. The arrangement of terms was developed by identifying 17 categories of drug-induced diseases and subdividing these into sometimes overlapping subcategories of more specific types of dysfunction or disease (see Table 13.3.2). Note that some of the definitions in Index E are broad. Therefore, although there is a high level of sensitivity, the false positive rate is high.

Since many companies use the WHOART and the FDA encourages the sponsors to also give the corresponding COSTART terms when reporting adverse drug reactions, there is a need to translate WHOART to COSTART. Indexes F and G describe the relationship between the COSTART dictionary and the dictionary adopted by the World Health Organization Adverse Reaction Terminology (WHOART). Index F of the COSTART dictionary lists the translation of each and every COSTART term to an acceptable WHOART term, while Index G of the COSTART dictionary provides translations of each and every preferred WHOART term to an acceptable COSTART term.

Table 13.3.2 General Categories of Drug-Induced Diseases

Category	Description
1	Automatic nervous
2	Cardiovascular
3	Endocrine
4	Gastrointestinal
5	Genitourinary
6	Gynecologic
7	Hematologic
8	Maternal-fetal
9	Metabolic
10	Nervous
11	General/nonspecific
12	Ophthalmic
13	Pathological
14	Pulmonary
15	Renal
16	Reticuloendothelial
17	Skin

The COSTART dictionary is widely used for the purpose of grouping similar events. Any potential trend in adverse events can be detected in this efficient way. However, there are potential problems with any dictionary used for grouping. First, we may map synonymous clinical events to different class terms. This is known as the problem of *one-to-many* problems. For example, *Heart attack* could be classified to either *Cardiovascular disorder*, *A-systole*, *Myocardial infarction*, or *Coronary artery disorder*. On the other hand, we may map clinically different events to a single class term, which is usually referred to as *many-to-one*. For example, the terms such as *Angiography*, *Coronary*, *Left ventricular aneurysm and thrombus*, and *Heart murmur* may be classified to the category of *Surgical procedure*. In practice, it is difficult to avoid these mapping problems.

13.3.3 MedDRA

As mentioned above, currently there is no uniform definition for adverse events. As a result, no internationally accepted medical terminology exists for evaluation of safety information for regulatory purpose. However, most pharmaceutical companies as well as regulatory agencies employ one of the international adverse drug reaction terminologies in combination with morbidity terminology. For example, regulatory agencies in Europe use a combination of WHOART and ICD-9th Revision (ICD-9). The COSTART adopted by the U.S. FDA is usually used in conjunction with ICD-9-CM (a clinical modification of ICD-9). On the other hand, the Japanese have developed their own version of the international terminologies, namely, J-ART and MEDIS. These established international medical terminologies and coding systems have been criticized for the lack of specificity of terms provided at the data entry level, limited data retrieval capability, and inability to handle syndromes effectively. As a result, regulatory agencies and pharmaceutical companies have used these international terminologies in a modified format or have developed their own *in-house* terminologies to overcome these deficiencies.

For a particular drug product, different terminologies and coding systems might be used at different stages of clinical development. For example, safety data such as an adverse event are frequently classified using ICD for preregistration clinical trials, whereas J-ART, WHOART, or COSTART are employed for postmarketing surveillance. This often makes it very difficult to cross-reference adverse event data across different times during the life span of the drug product. In addition, different regions or countries may use different terminologies or coding systems for the same adverse events. This problem exacerbates when different subsidiaries of global pharmaceutical companies use different terminologies because of variations in regulatory requirements for submission of adverse data. This problem, however, has been recognized by the regulatory agencies as well as by the pharmaceutical industry for a long time. The resolution of this problem requires standardization of the content and structure of the data set.

In 1989, the U.K. Medical Control Agency (MCA) identified the need for a single medical terminology to support classification of terms relating to all aspects of drug regulations for use within the MCA computer database. The MCA medical terminology was introduced in the MCA pharmaco-vigilance database in 1991. In 1993, the MedDRA Working Party was formed to assess the potential applicability of MCA medical terminology for other regulatory agencies and the pharmaceutical industry. At the end of 1993, the CPMP approved MedDRA remit and objectives. MedDRA version 1.0 was completed in August 1994. The ICH Steering Committee adopts MedDRA version 1.0 as the basis for the new international medical terminology for regulatory purposes. In September 2002, MedDRA version 5.1 was released (MedDRA Introductory Guide, 2002). Most regulatory agencies in the ICH region have subscribed to the MedDRA.

The MedDRA terminology is intended to apply to all phases of drug product (biologicals) development, excluding animal toxicology. It is also applied to the health effects of devices (e.g., uteric rupture after intrauterine device insertion, or infection because of failure to sterilize a catheter). To achieve the above objectives, the terms in MedDRA are classified in the following categories:

- Symptoms
- Signs
- Diseases
- Diagnosis
- Therapeutic indications
- Names and qualitative results of investigations, including pharmacokinetics
- Surgical and medical procedures
- Medical/social/family history

In addition, the terminology defined above has been developed for regulatory authorities and the regulated medical product industry for utilization in data entry, retrieval, evaluation, and presentation, in both pre- and postmarketing phases of the regulatory process as follows:

- Clinical trials
- Reports of spontaneous adverse reactions and events
- Regulatory submission
- Regulatory product information

Table 13.3.3 Inclusion of Terms from Established Terminologies by MedDRA

WHO-ART (98:3)	Preferred Terms, Included Terms
J-ART (1996)	Preferred Terms, Included Terms
COSTART (Fifth Edition)	Preferred Terms, Glossary Terms
ICD-9	Terms associated with 3 and 4 digits
ICD-9-CM (Fourth Edition)	Terms associated with 3,4 and 5 digits
HARTS (Release 2.2)	

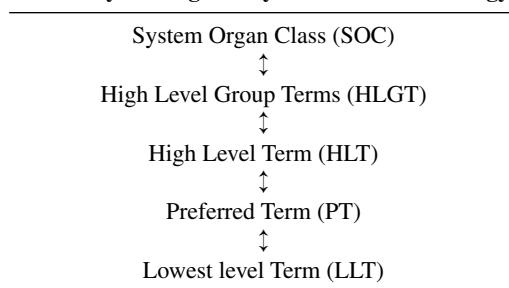
Table 13.3.3 lists the established terminologies included in the MedDRA terminology. However, the MedDRA has not been developed as a metathesaurus and other established terminologies such as COSTART or WHOART are not subsets of MedDRA. Therefore, data entry terms from other terminologies do not necessarily have the same Preferred Term (PT) in MedDRA terminology. Basically, the MedDRA terminology comprises medical terms with their unique codes organized and grouped in a hierarchical and horizontal fashion through three categories.

Equivalence A term in the MedDRA terminology is a single medical concept that may be reported by investigators using different but equivalent terms such as verbatim terms or reported terms. For example, *fatigue* is an underlying medical concept that can be reported as exhaustion, loss of energy, tiredness, worn out, tire out, and so on. The MedDRA groups synonymous terms or terms regarded as equivalent under one standard term called Preferred Term (PT).

Hierarchical These Preferred Terms are grouped under a five-level hierarchical framework of medically meaningful terms. This hierarchy provides levels of superordination or subordination, in which the superordinate term is a broad grouping term applicable to each subordinate descriptor linked to it. As a result, this hierarchical framework represents vertical links in the MedDRA terminology and provides an important mechanism for flexible data retrieval and presentation of data for easy review. Table 13.3.4 provides the structure hierarchy of the MedDRA terminology. At the bottom of the hierarchical is the Lowest Level Terms (LLT), which are equivalent terms for a single medical concept. Each LLT is linked to a PT in any of the following relationships:

- Synonyms—different terms for the same descriptor
- Lexical variant—different word forms for the same expression (e.g., full names versus abbreviations) and direct versus inverted word order included from preexisting terminologies
- Quasi-synonyms—terms where meanings are generally regarded as different, but that can be treated as equivalent clinically for medical product regulatory purposes

For the purpose of data entry, one LLT is identical to the PT. Preferred Terms in the MedDRA terminology is a distinct descriptor (i.e., a single medical concept) for a symptom, sign, disease, diagnosis, therapeutic indication, investigation, surgical or medical procedure, medical, social, or family history characteristic. PT is used to group equivalent LLTs. As a result, when there are one or more equivalent terms, the term selected as PT is preferred for use in the regulatory environment, formatted according to the terminology

Table 13.3.4 Structure Hierarchy of the Medical Dictionary for Regulatory Activities Terminology

conventions. In other words, in the context of regulatory and international requirements, PTs should be unambiguous and as specific and self-descriptive as possible. In addition, the granularity/specificity of the PT level is such that clinical pathologic or etiological qualifiers of the descriptors are represented at the PT level. This level of specificity ensures that the multiaxial nature of the MedDRA terminology can be exploited maximally. There is no limit to the number of LLTs that can be linked to a PT. However, a PT term must have at least one LLT below it.

The next two higher levels in the hierarchy are the High Level Terms (HLT) and High Level Group Terms (HLGT). An HLT is a superordinate descriptor for the PTs linked to it inclusively by anatomy, pathology, physiology, etiology, or functions. Because the terminology is not taxonomy, the specificity of HTLs is not uniform throughout the MedDRA terminology. An HLT is a superordinate descriptor for one or more HTLs, related also by anatomy, pathology, physiology, etiology, or functions. However, HLT groups HTLs together to aid retrieval by broader concepts. HLTs are subordinate to System Organ Class (SOC), which is the highest level of the hierarchy and provides the broadest concept for data retrieval. SOCs comprise grouping by

- Anatomical or physiological system (e.g., gastrointestinal disorders)
- Body organ (e.g., eye disorders)
- Etiology (e.g., infections and infestations)
- Purpose (e.g., surgical and medical procedures)

There are a total of 26 SOCs, which are listed in Table 13.3.5 in an internationally agreed order. Each term in the MedDRA terminology is assigned a unique nonexpressive eight-digit code. For example, *Gastric Cancer* is a single underlying medical concept. Therefore, *Gastric Cancer* is a PT with code 10017760 in MedDRA with a total of 24 LLTs below it, including stomach cancer (10017758), stomach carcinoma (10042090), carcinoma, stomach (10007474), gastric carcinoma (1001770), gastric malignancy (10017796), carcinoma, gastric (10007350), malignant stomach neoplasm, malignant stomach tumor, malignant gastric neoplasm, and gastric cancer. The PT *Gastric Cancer* (10017760) is linked to SOC *Neoplasm benign and malignant (including cysts and polyps)* (10029104) through the route via HLT *Gastric neoplasms malignant* (10017812) and HLT *Gastrointestinal neoplasms and malignant and unspecified* (10017991). The numbers in the parentheses are nonexpressive eight-digit codes from which no information can be derived. In fact, these eight-digit numeric codes are applied to all terms across all categories. Initially, the codes

Table 13.3.5 The Medical Dictionary for Regulatory Activities Terminology System Organ Class Internationally Agreed Order

Infection and infestations
Neoplasms benign and malignant (including cysts and polyps)
Blood and lymphatic system disorders
Immune system disorders
Endocrine disorders
Metabolism and nutrition disorders
Psychiatric disorders
Nervous system disorders
Eye disorders
Ear and labyrinth disorders
Cardiac disorders
Vascular disorders
Respiratory, thoracic and mediastinal disorders
Gastrointestinal disorders
Hepato-biliary disorders
Skin and subcutaneous tissue disorders
Musculoskeletal, connective tissue and bone disorders
Renal and urinary disorders
Pregnancy, puerperium and perinatal conditions
Reproductive system and breast disorders
Congenital and familial/genetic disorders
General disorders and administration site disorders
Investigations
Injury and poisoning
Surgical and medical procedures
Social circumstances

were assigned in alphabetical order starting with 10000001. Codes will not be reused, even if a term is not used anymore. New terms added to the MedDRA terminology will be assigned the next available sequential number. Currently, there are a total of 26 SOCs, 334 HLTGs, 1,663 HLTs, approximately 16,100 PTs, and 57,000 LLTs.

From the above example, a PT must be linked to at least one SOC via an HLT and one HLTG. However, it can only be linked to a particular one SOC via one route. In addition, an HLT must be also linked to at least one SOC and an HLTG must be linked to at least one SOC and one HLT. There is no limit to the number of SOCs to which an HLTG can be linked. As a result, PTs can be represented in more than one SOC. To overcome the problem of double counting, PTs in the MedDRA terminology are assigned a Primary System Organ Class, which determines the SOC where the term is displayed in these outputs. This facility does not prevent display and counting of the term in any of the SOCs in which it is represented for data retrieval purposes, which do not involve all SOCs.

Special Search Categories (SSC) Special Search categories link groups of PTs horizontally via an associative relationship in the MedDRA terminology. In general, SSCs are used to group PTs relevant to an issue, usually a disease or syndrome, and accommodate clinical concepts that cross SOC hierarchies. Table 13.3.6 lists the nine Special Search Categories in MedDRA terminology. For example, hemorrhage terms are scattered throughout the body system and organ SOCs, but they form a useful data retrieval grouping. However, the

Table 13.3.6 The Medical Dictionary for Regulatory Activities Terminology Special Search Categories

Anaphylaxis
Arrest (Cardiac)
Blood Dyscrasias/Bone marrow depression
Cardiac Ischaemia
Hemorrhage
Hypersensitivity reactions
Thrombosis
Upper GI bleeding/perforation
Vasculitis

hierarchy of the blood and the lymphatic system disorders' SOC reflects established classification groupings, which precludes retrieval, via the HLGT/HLT hierarchy, of all PTs relevant to bone marrow suppression, an important grouping in medical product regulatory terms. An SSC blood dyscrasias/bone marrow depression was formed to overcome this deficiency. There is no restriction to the number of PTs that can be linked to SCC, and terms for any SOC may be included; SSCs usually comprise PTs belonging to different hierarchies, but they may include terms from the same hierarchy.

13.3.4 Common Toxicity Criteria

As indicated above, because of the life-threatening nature of cancer and the severity and complexity of toxicity associated with most cancer treatments, both the U.S. NCI and EORTC recommend the use of the Common Toxicity Criteria (CTC Manual, version 2.0, 1999) to evaluate adverse events for their sponsored cancer clinical trials. The primary organization of the CTC is based on path physiological and anatomical categorical to facilitate location of related adverse events. Table 13.3.7 provides a list of these 24 categories. Within each of these categories, specific adverse events are listed alphabetically with the grade of severity. There are a total of 200 individual adverse events, which are also classified by occurrence with use of investigational treatment, including chemotherapy, biological therapy, radiation therapy, and surgery. In addition, for selected adverse events, different grading criteria are provided together for different patient populations. In other words, the adverse event name is the same, but the criteria for grading are changed.

For each adverse event, a grade is assigned and defined using a scale from 0 to 5, with 0 representing no adverse event within normal limits and 5 representing death related to an adverse event:

- 0 = No adverse event or within normal limits
- 1 = Mild adverse event
- 2 = Moderate adverse event
- 3 = Severe and undesirable adverse event
- 4 = Life-threatening or disabling adverse event
- 5 = Death related to adverse event

One of the most difficult parts in evaluating and grading an adverse event is to distinguish a treatment-related condition from a disease-related condition. CTC v2.0 specifies that

Table 13.3.7 Categories in the Common Toxicity Criteria

Allergy/Immunology	Infection or Febrile Neutropenia
Auditory/Hearing	Lymphatics
Blood/Bone marrow	Metabolic/Laboratory
Cardiovascular (Arrhythmia)	Musculoskeletal
Cardiovascular (General)	Neurology
Coagulation	Ocular/Visual
Constitutional Symptoms	Pain
Dermatology/Skin	Pulmonary
Endocrine	Renal/Genitourinary
Gastrointestinal	Secondary malignancies
Hemorrhage	Sexual/Reproductive function
Hepatic	Syndromes

disease progression signs and symptoms definitely related to disease should not be graded. In addition, treatment delivery system malfunctions or sequelae only to the treatment delivery system, such as a broken needle requiring excision, should not be graded as adverse events. All other symptoms, signs, or diseases (including abnormal laboratory findings) that might be associated with investigational drugs or therapies must be captured and graded. Table 13.3.8 gives the system for evaluation of attribution of causality of adverse events used by the CTC v2.0. If an adverse event is of any probable, possible, or definite relationship to the investigational drugs, CTC v2.0 recommends that investigators must document and grade the adverse event. However, adverse events that are definitely not related to disease should not be graded. On the other hand, if an adverse event is caused by a combination of treatment and disease, the adverse event should be graded as it is observed with no adjustment. Furthermore, the CTC v2.0 requires that for each event, the physician or clinician in conjunction with the research nurse who examined and evaluated the patient should assign the attribution. The CTC v2.0 Manual specifically recommends that the assignment of the attribution for adverse events should not be performed by the data managers who have been removed from the clinical assessment of the patient.

In 1999, as part of its commitment to the ICH, the U.S. FDA agreed to adopt an internationally agreed International Medical Terminology (IMT) based on the U.K. MCA MedDRA for use in reporting medical information from clinical trials. To facilitate data transfer, the U.S. NCI has supported the mapping of adverse event names from the CTC v2.0 to PT in the IMT. However, as a result of international harmonization, currently, all

Table 13.3.8 Attribution of Adverse Events by the Common Toxicity Criteria

Code	Descriptor	Definition
5	Definite	The adverse event is clearly related to the investigator agents
4	Probable	The adverse event is likely related to the investigator agents
3	Possible	The adverse event may be related to the investigator agents
2	Unlikely	The adverse event is doubtfully related to the investigator agents
1	Unrelated	The adverse event is clearly not related to the investigator agents

references to formally IMT Codes and Terms have been replaced by the MedDRA terminology described in Section 13.3.3. The NCI has established a computer data management system called “Clinical Data Update System version 3” (CDUS version 3.0, 2002) to collect adverse events for its sponsored cancer clinical trials. In addition, it also issued guidelines on Expedited Adverse Event Reporting Requirements for NCI Investigational Agents (2001) and established an electronic system, the Adverse Event Expedited Reporting System (AdEERS), for expedited submission of adverse events, including serious and unexpected events. For more details of CTC v2.0, CDUS version 3.0, and AdEERS, visit the CTEP homepage at <http://ctep.info.nih.gov>.

Some Concerns

How should syndromes be recorded on the case report form? as the syndrome, as symptom separately, or both? A syndrome is a collection of signs, symptoms, or laboratory findings that together describes a distinct clinical entity. For example, congestive heart failure is described by one or more of shortness of breath, peripheral edema, easy fatigability, pulmonary rales, elevated venous pressures, low plasma sodium, and elevated BUN. The commonly encountered problem with the summarizing syndrome and its components is that symptoms that occur alone are not distinguished from those that occur as part of the syndrome. Moreover, increased incidence of true adverse events related to the drug may be lost in a high incidence of the same event that occurs as a component of a syndrome in both groups. These problems may misrepresent the effect of the drug.

How should primary versus secondary adverse event be recorded? The problems include (1) multiple counting of same events and (2) over-reporting of components of syndrome or secondary events, which are not in and of themselves related to drug. These concerns may lead to inaccurate conclusions about undesirable effects of the drug. Thus, analysis of primary events leads to correct understanding of causal relationship between the drug’s administration and adverse events. Both primary and secondary adverse events can be recorded on CRF, with the secondary event clearly linked to the primary event or syndrome.

13.4 ANALYSIS OF ADVERSE EVENTS

As indicated earlier, the evaluation of safety-related data can be considered at three levels. The next levels involve the more common adverse events and serious adverse events. The analysis of adverse events is not only to identify factors that may affect the frequency of adverse reactions/events such as time dependence, relation to demographic characteristics, relation to dose or drug concentration but also to determine whether the identified adverse events are drug related. For this purpose, we may consider the approaches of data listing, summary tables, and statistical analysis.

Adverse Event Data Listing

For evaluation of adverse events, it is helpful to list all adverse events including the same event on several occasions by the patient and by the treatment group. The listing should include both the preferred term and the original term used by the investigators. For a complete listing of adverse events, the ICH guidelines suggest that the information listed in Table 13.4.1 be included in the data listing.

Table 13.4.1 Information to be Included in the List of Adverse Events

Patient identifier
Age, race, sex, weight
Location of case report forms if provided
Adverse event
Duration of the adverse event
Severity
Seriousness
Action taken
Outcome
Causality assessment
Date of onset or date of clinic visit at which the event was discovered
Timing of onset of the adverse event in relation to the last dose of the last drug
Study treatment at the time of event or the most recent study taken
Test drug/investigational product dose in absolute amount, mg/kg or mg/m ² , at time of event
Drug concentration (if known)
Duration of test drug/investigational product treatment
Concomitant treatment during study

In Table 13.4.1 the patient characteristics may include height if relevant. The adverse event should be displayed both as the preferred term and the reported term. The degree of severity may be classified as mild, moderate, or severe, while the seriousness can be expressed as either serious or nonserious. For action taken, it could be none, dose reduced, treatment stopped, specific treatment instituted, and so forth. The causality assessment may include two categories of related and not related. However, how it was determined should be clearly described in the protocol.

Note that also required is that all deaths during the study, including the post-treatment follow-up period and deaths that resulted from a process that began during the study should be listed by patient. All serious adverse events other than death but including the serious adverse events temporally associated with or preceding the deaths should be listed. In addition other significant adverse events and any events that led to an intervention, including withdrawal of test drug/investigational product treatment, dose reduction, or significant additional concomitant therapy, other than those adverse events reported as serious should be listed.

Summary Tables of Adverse Events

The ICH guidelines suggest that all adverse events occurring after the initiation of study treatments be displayed in summary tables. The adverse events include events that are likely to be related to the underlying disease or events that are likely to represent a concomitant illness. The summary tables list each adverse event, the number of patients in each treatment group in whom the event occurred, and the rate of occurrence. In most cases it is helpful to identify and summarize events not seen at the baseline and events that worsened even if present at the baseline. Moreover, the summary tables should include changes in vital signs and any laboratory changes that are considered serious adverse events or other significant adverse events.

In practice, adverse events are grouped by body system and then divided into defined severity categories. However, in presenting adverse events, it is important to display both the original terms used by the investigators and to attempt to group related events so that the true occurrence rate is not obscured. The ICH guidelines recommend that a standard adverse reaction/events dictionary be used for grouping. As indicated in the previous section, the ICH MedDRA and the COSTART are probably the most commonly used dictionary for this purpose.

In most cases the summary tables help divide the adverse events into those considered at least possibly related to drug use and those considered unrelated, or by another causality scheme such as unrelated or possibly, probably, or definitely related. However, when a causality scheme is used, it is suggested that the summary tables include all adverse events regardless of whether they are drug related or represent intercurrent illnesses. It is also important to identify each patient having the adverse event for safety evaluation. An example of such a tabular presentation as specified in the ICH guidelines is reproduced in Table 13.4.2. In addition to Table 13.4.2, the summary tables should include a comparison of common adverse events between the treatment and control groups. The observed safety profiles from the summary tables can then be confirmed by some formal statistical tests which will be discussed in the next subsection.

Graphical Presentation

To provide a preliminary examination of adverse events between treatment groups, a graphical presentation is usually helpful. A creative graphical presentation enables clinical scientists/researchers to discover trends/patterns about the data that are unanticipated. Levine (1996) recommends that a study design and patient flow display, raw data display, and inferential displays of clinical trial adverse events be provided. Levine indicates that these graphical presentations provide not only rapid answers to prespecified questions but also the insight into the structure of the raw data. In addition a creative graphical presentation can generate new questions regarding safety and provide rapid assessment to these new questions. For example, Figure 13.4.1 gives a scatter plot showing adverse events by the body system against the percentage of patients reporting events from a placebo-controlled clinical trial. This scatter plot not only provides a preliminary examination of adverse events rates within treatments but also compares the adverse event rates between treatments.

Analysis of Adverse Events

Although adverse event data listing, summary tables of adverse events, and graphical presentations of adverse events provide useful information for the safety evaluation of a study drug, they do not provide any statistical inference for the safety assessment. As was indicated earlier, most clinical studies are not designed to examine safety. As a result the analysis of adverse events is less standardized than that for efficacy data. Silliman (1996) classifies adverse events into rare adverse events and common adverse events. For rare adverse events the primary interest is to estimate the occurrence. For common adverse events, in addition to estimation of occurrence, it is also of interest to make comparisons between treatments and/or among subgroups. Basically the analysis of adverse events depends on the type of data, which can be classified as nominal (binary) or ordinal, counts or rates, or time to occurrence. These types of data are among those described in Section 13.2 and are usually analyzed by Fisher's exact test or the Mantel-Haenszel test,

Table 13.4.2 Adverse Events: Number Observed and Rate with Patient Identifications

	Treatment Group X						<i>N</i> =50	
	Mild		Moderate		Severe			
	Related ^a	NR ^a	Related	NR	Related	NR		
Body System A								
Event 1	6(12%) <i>N11^b</i>	2(4%) <i>N21</i>	3(6%) <i>N31</i>	1(2%) <i>N41</i>	3(6%) <i>N51</i>	1(2%) <i>N61</i>	12(24%) 4(8%)	
	<i>N12</i>	<i>N22</i>	<i>N32</i>	<i>N33</i>	<i>N52</i>	<i>N53</i>		
	<i>N13</i>							
	<i>N14</i>							
	<i>N15</i>							
	<i>N16</i>							

Event 2

^aNR = not related; related could be expanded (e.g., as definite, probable, possible).^bPatient identification number.

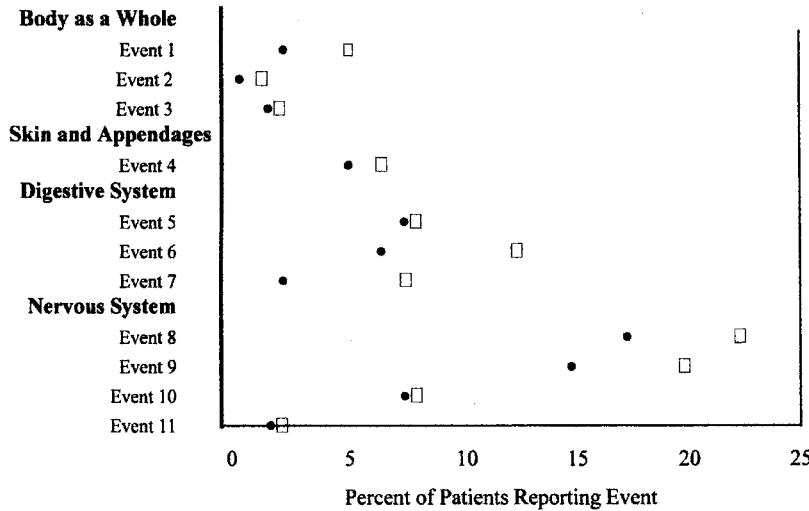


Figure 13.4.1 Scatter plot of adverse events versus incidence rates: ● placebo, □ drug.

logistic regression, and survival analysis. In this section, for illustration purposes, we will only consider the commonly used approaches for comparison of adverse event rates between treatments (e.g., treatment vs. control). For other types of adverse events data such as absorbing events and recurring events with and/or without duration as described in Section 13.2, statistical methods including the Kaplan-Meier and the Cox proportional hazards models, as described in Chapter 10, can be directly applied.

For simplicity, consider the case where the event severity categories and/or the causality categories of a given response (adverse event) are combined into a 2×2 contingency table as shown in Table 13.4.3. For a 2×2 contingency table, a typical approach is to use the Mantel-Haenszel test. For the 2×2 table given in Table 13.4.3, the Mantel-Haenszel test is given by

$$\chi^2_{\text{MH}} = \frac{N(ad - bc)^2}{n_1 n_2 m_1 m_2} \quad (13.4.1)$$

or with a continuity correction as

$$\chi^2_{\text{MH}} = \frac{N(|ad - bc| - n/2)^2}{n_1 n_2 m_1 m_2}$$

Table 13.4.3 The 2×2 Contingency Table for Adverse Events

	Adverse Events		Total
	Presence	Absence	
Treatment	a	b	n_1
Control	c	d	n_2
Total	m_1	m_2	N

Note that χ^2_{MH} can be written as

$$\begin{aligned}\chi^2_{\text{MH}} &= \frac{N}{N-1} \left[\frac{a-E(a)}{\sqrt{\text{var}(a)}} \right]^2 \\ &= \frac{N}{N-1} \chi^2_p,\end{aligned}$$

where

$$E(a) = \frac{n_1 m_1}{N}$$

and

$$\text{var}(a) = \frac{n_1 n_2 m_1 m_2}{N^2(N-1)}.$$

When the sample size is sufficiently large, the Mantel-Haenszel test is approximately distributed as a chi-square random variable with one degree of freedom. Therefore, we can reject the null hypothesis of no difference between treatment groups at the α level of significance if

$$\chi^2_{\text{MH}} > \chi^2(\alpha, 1).$$

When the sample size is small, Fisher's exact test as described in (9.3.16) is often used as an alternative to the Mantel-Haenszel test. For a 2×2 table, Fisher's exact test yields the probability of observing a table that gives at least as much evidence of association as the one actually observed, given that the null hypothesis is true. In other words, with row and column margins, namely n_1, n_2, m_1, m_2 considered fixed, the distribution of a has the hypergeometric form

$$P(a|n_1, n_2, m_1, m_2) = \frac{\binom{n_1}{a} \binom{n_2}{m_1-a}}{\binom{N}{m_1}},$$

where a takes on values between $\max(0, n_1 + m_1 - M)$ and $M = \min(n_1, m_1)$. Thus we can reject the null hypothesis of no difference if

$$p\text{-value} = \sum_{i=1}^M P(i|n_1, n_2, m_1, m_2) \quad (13.4.2)$$

is less than the α level of significance.

In clinical trials, although an adverse event can be classified into dichotomous groups such as presence versus absence or serious versus nonserious (see Table 13.4.3), it is often of interest to examine the intensity or severity of an adverse event that may be evaluated according to c categories such as *mild*, *moderate*, and *severe* (where $c = 3$). In this case, the 2×2 table becomes $2 \times c$ table. On the other hand, the intended clinical trial may be designed to compare r treatments. In this case the 2×2 table becomes a $r \times c$ table. Both the Mantel-Haenszel test and Fisher's exact test can be extended to analyze $r \times c$ tables. Note that the Mantel-Haenszel test for $r \times c$ tables is also known as the Cochran-Mantel-Haenszel test. Agresti and Wackerly (1977) consider generations of Fisher's exact test for $r \times c$ tables. Pagano and Halvorsen (1981) propose an algorithm, which does not require the total

enumeration of all tables consistent with the given marginals, for calculating the exact permutation significance value for $r \times c$ tables.

In practice, since most clinical trials are conducted at more than one study site, it is necessary to adjust for the effects that may be due to the study site (or center) when comparing adverse event rates between treatment groups. For this purpose the Cochran-Mantel-Haenszel (CMH) test is useful. The concept behind the CMH test is to consider the study site (or center) as a stratum and assume that the strata are independent and that the marginal totals of each stratum (center) are fixed. Under these assumptions the null hypothesis is that there is no association between treatment and response (a given adverse event say) in any of the strata. Consequently the multiple hypergeometric model can be used to derive the CMH test. Let $\mathbf{n}_{hi} = (n_{hi1}, n_{hi2}, \dots, n_{hic})'$, where n_{hij} is the cell frequency of the response at the j th intensity or severity category for the i th treatment group at the h th center, $i = 1, \dots, r$ (the number of treatments), $j = 1, \dots, c$ (the number of categories for the response), and $h = 1, \dots, H$ (the number of centers). Also let

$$\mathbf{n}_h = (\mathbf{n}_{h1}, \mathbf{n}_{h2}, \dots, \mathbf{n}_{hr})'$$

and

$$P_{hi.} = \frac{n_{hi.}}{N_h}, \quad P_{hj} = \frac{n_{hj.}}{N_h},$$

where N_h is the total at the h th center. Thus we have

$$\mathbf{P}_{h*} = (P_{h1.}, P_{h2.}, \dots, P_{hr.})'$$

and

$$\mathbf{P}_{h.*} = (P_{h.1}, P_{h.2}, \dots, P_{h.c}).$$

Under the null hypothesis the expected value and covariance matrix of the frequencies are then given by

$$\mathbf{m}_h = E(\mathbf{n}_h) = N_h(\mathbf{P}_{h.*} \otimes \mathbf{P}_{h.*})$$

and

$$\text{var}(\mathbf{n}_h) = \frac{N_h^2}{N_h - 1} [(\mathbf{D}_{P_{h.*}} - \mathbf{P}_{h.*}\mathbf{P}'_{h.*}) \otimes (\mathbf{D}_{P_{h.*}} - \mathbf{P}_{h.*}\mathbf{P}'_{h.*})],$$

where \otimes denotes Kronecker product multiplication and \mathbf{D}_A is a diagonal matrix with elements of A on the main diagonal. The generalized CMH test is then defined as

$$\chi^2_{\text{CMH}} = \mathbf{G}' \mathbf{V}_G^{-1} \mathbf{G}, \quad (13.4.3)$$

where

$$\mathbf{G} = \sum_h \mathbf{B}_h (\mathbf{n}_h - \mathbf{m}_h),$$

$$\mathbf{V}_G = \sum_h \mathbf{B}_h \text{var}(\mathbf{n}_h) \mathbf{B}'_h,$$

and where

$$\mathbf{B}_h = \mathbf{C}_h \otimes \mathbf{R}_h$$

is a matrix of fixed constants based on column scores \mathbf{C}_h and row scores \mathbf{R}_h . Under the null hypothesis, χ^2_{CMH} is approximately distributed as a chi-square random variable with degrees of freedom equal to the rank of \mathbf{B}_h . Note that the CMH test given in (13.4.3) is for the general case of $r \times c$ tables stratifying for the study sites. When $r = c = 2$, the CMH test can be expressed as

$$\chi^2_{\text{CMH}} = \frac{\left(\sum_{h=1}^H [a_h - (n_{1h}m_{1h}/N_h)] \right)^2}{\sum_{h=1}^H [n_{1h}n_{2h}m_{1h}m_{2h}/N_h^2(N_h - 1)]},$$

where a_h , n_{1h} , n_{2h} , m_{1h} , m_{2h} , and N_h are defined as in Table 13.4.3 for the h th center.

As was indicated earlier, for a premarketing safety assessment, the FDA requires that an integrated summary of all available information about the safety of the drug product be submitted (Section 314 of CFR [314.50 (d)(5)(ii)]). The purpose of such an integrated summary is to allow examination of differences among population subsets not possible with the relatively small numbers of patients in individual studies. Therefore the integrated summary is, in part, simply a summarization of data from individual studies and, in part, a new analysis that goes beyond what can be done with individual studies. Silliman (1996) suggests that combining data across studies in the integrated summary of safety be done by either subgroup analysis (pooling the data) or by performing a meta-analysis. For analysis of subgroup safety data by pooling, the Fisher exact test, the CMH test (or Breslow-Day test), logistic regression, and survival analysis are commonly employed. The objectives of subgroups analysis are to address the following questions in an integrated summary safety report:

1. Are adverse event rates the same across a subgroup for patients taking experimental drug?
2. Within subgroup levels, are adverse event rates the same across treatment groups?
3. Is there a consistent association between the treatment group and the adverse event response across levels of a subgroup?
4. Does the subgroup predict an adverse event response?
5. Is the time the occurrence of the adverse event the same across levels of a subgroup?

For an analysis of subgroup safety data using meta-analysis, the commonly used statistical methods include the CMH test (or Breslow-Day test), logistic regression such as proportional odds model, and survival analysis such as the Kaplan-Meier (log-rank), piecewise exponential, and incidence density tests. The objective of a meta-analysis is to address the above questions by controlling for the study and/or time interval.

13.5 ANALYSIS OF LABORATORY DATA

In clinical trials the analysis plan for laboratory data is usually vague. A typical approach is to compare the change from baseline to the last patient visit both between and within each treatment group. The comparison can be made either based on the absolute mean

change from baseline in laboratory values or based on the frequencies of the values outside the reference or normal ranges. Either way provides a different analysis and interpretation of the laboratory data. In this section we will focus on the analysis of frequencies of the values according to the reference or normal ranges. For analysis of the absolute mean changes from baseline, the statistical methods described in Chapter 8 can be used.

Reference Ranges

In practice, most clinical trials are conducted at more than one study site in order to enroll enough patients within a desired time frame. In this case a concern may be whether the laboratory tests should be performed by local laboratories or a central laboratory. The relative advantages and drawbacks between the use of a central laboratory and local laboratories include (1) the combinability of data, (2) timely access to laboratory data, (3) laboratory data management, and (4) cost. For example, a central laboratory provides combinable data with unique normal ranges, while local laboratories may produce uncombinable data due to different equipment, analysts, and normal ranges. As a result laboratory data obtained from central laboratories are more accurate and reliable compared with those obtained from local laboratories.

Combinability of Laboratory Data

Since different local laboratories may have different normal ranges, it is necessary to transfer laboratory results according to the investigators' normal ranges or local laboratories' normal ranges to a *standard* range before analysis. Typically, a laboratory result can be transformed according to the following formula:

$$R_t = S_L + \frac{R_u - I_L}{I_H - I_L} (S_H - S_L). \quad (13.5.1)$$

where R_u and R_t denote the untransformed result and transformed result, respectively, and (I_L, I_H) and (S_L, S_H) are the investigators' and *standard* lower and upper limits of normality, respectively. Note that if both lower limits are equal to 0, then the transformation is based only on the ratio of upper limits. Moreover, if an untransformed result is within the investigator's limits, then the transformed result will be within the *standard* limits. However, data cannot be transformed when lower and upper limits are equal. An example of such a case is hemoglobinuria whose normal ranges are known to be $(0, 0)$; that is, complete absence is the only normal condition. As can be seen from (13.5.1), if results for a specific test have different investigators' ranges, the transformation procedure effectively removes data variations due to the different sources to be examined under equivalent conditions and displays comparable values.

Chuang-Stein (1996) propose similar methods to combine data from different laboratories. Let I_W be the width of the investigators' normal range,

$$I_W = I_H - I_L.$$

Then

$$I_M = \frac{1}{2} (I_L + I_H)$$

and

$$I_S = \frac{1}{2} (I_H - I_M)$$

are the midpoint and the $\frac{1}{4}$ of the investigators' normal range. Chuang-Stein (1996) suggests using the following two methods for combining data from different laboratories: The first method is to combine the data by means of the transformation

$$R_1 = \frac{R_u - I_L}{I_W}. \quad (13.5.2)$$

As can be seen from the above, $R_1 \in (0,1)$ when $R_u \in (I_L, I_H)$; otherwise, R_1 is outside of the interval. The second method is to consider

$$R_2 = \frac{R_u - I_M}{I_S}. \quad (13.5.3)$$

When $R_u \in (I_L, I_H)$, R_2 is within the interval of $(-2, 2)$; otherwise, it is outside of this interval.

For safety assessment based on laboratory data, correct normal ranges are important for the following reasons: First, notable outlying values and/or clinically significant changes from the baseline must be identified in order to protect the patients during the clinical trial. Second, patients who exhibit unusual laboratory results may be excluded from the trial. In addition, patients with notable test results must be identified, studied, and discussed in the final study report. As a result the determination of normal ranges will have an impact on the assessment of safety of the drug under study. In practice, the commonly encountered difficulties with normal ranges are (1) normal ranges are usually based on healthy volunteers who may not be representative of the patient population of the intended clinical trial, (2) some normal ranges may be inaccurate due to a small-sample estimate, and (3) the underlying distribution of many laboratory analyses is not normally distributed. To overcome these difficulties, it is suggested that the so-called Lilly reference limits be used. The concept behind Lilly reference limits is to use a large number of laboratory data obtained from the baseline visits. The laboratory data are first examined for the effect of various covariates such as age, gender, origin, smoking, and alcohol use. Lilly reference limits are then obtained using actual percentiles without any distributional assumptions.

Note that before the data from different laboratories can be combined for analysis, it may be of interest to evaluate the repeatability and reproducibility of the results. The repeatability of the results is referred to as within-laboratory variability, while the reproducibility of the results is the assessment of between-site variability. Typically the repeatability and reproducibility can be assessed by sending to each laboratory identical samples that represent a wide range of possible values. The results can then be analyzed using the method analysis of variance.

Clinically Significant Changes

Before one can determine whether there is a clinically significant change in laboratory data, the definition of clinically significant change is essential. The definition of clinically significant change depends on the objectives of the laboratory evaluation. For example, if the objective is to determine whether a patient's change in laboratory value from the baseline is within the normal range, then we can claim that a clinically significant change has

occurred provided that the patient's change in laboratory value from the baseline results in moving the patient from the status of normality to abnormality or from abnormality to normality. For assessing abnormal change, the most straightforward approach is to transform the result to the categories of either *normal* and *abnormal* according to some established reference or normal range.

The classification of patients' laboratory results to either *normal* or *abnormal* categories, however, cannot assess the magnitude of change. To assess the magnitude of change in terms of absolute (or actual) change or relative (percent) change, we need to establish an equivalence range so that any changes in laboratory values from the baseline within the range is not of clinical significance. This equivalence range can then be used as a reference range for determination of a clinically significant change. It should be noted that a clinically significant change may not result in a change in the status of normality or abnormality. To provide a consistent assessment, some pharmaceutical companies have suggested not to classify a subject into the category of *abnormal* unless (1) the result is outside of the normal range in a direction potentially adverse and (2) the change is of clinical significance. In practice, one may want to divide the magnitude of change from the baseline into a number of ranges to distinguish between different grades of change. For example, for a laboratory test of hematocrit, the normal range for a male is between 38% and 53% (see Table 13.2.2). We may consider three grades of change in either direction (i.e., increasing or decreasing direction) as 5% to 10% (grade 1), 10% to 15% (grade 2), and greater than 15% (grade 3). In this case clinically significant changes depend on the magnitude of change in the grades of interest.

In some clinical trials a patient's laboratory results are classified into several categories to assess the degree of concern depending on the frequency with which the abnormalities, singly or in succession, were found. Table 13.5.1 provides some examples of categories used in assessing the degree of concern. From Table 13.5.1 it can be seen that the categories range from all results normal (category I), to two or more successive abnormalities, and to the last observed value was abnormal (category V). That is to say, the definition of a clinically significant change is a change from one category to another category.

In many clinical trials, laboratory tests are performed at the visits (or time points). The change is assessed as a trend over time or at some specific time points. For patients without a value at a specified time, but with a value prior to and another following the specified

Table 13.5.1 An Example of Categories for Assessing the Degree of Concern

Category	Definition
I	All results normal
II	1. No two successive abnormalities 2. Last result normal
III	1. Two or more successive abnormalities 2. Last result normal
IV	1. No two successive abnormalities 2. Last result abnormal
V	1. Two or more successive abnormalities 2. Last result abnormal

time, an estimate at the specified time can be made by linear interpolation between given times as follows:

$$R_k = R_{i-1} \left(\frac{t_i - t_k}{t_i - t_{i-1}} \right) + R_i \left(\frac{t_k - t_{i-1}}{t_i - t_{i-1}} \right),$$

where R_{i-1} is the nearest result preceding the desired result (R_k), R_i is the nearest result following R_k , and t_k is the desired observation time. The definition of a clinically significant change for detection of a trend over time is based on the slope of the fitted linear regression if the relationship between the laboratory results and time is linear. For specific time points, a clinically significant change may depend on the disease status or duration of the treatment at that time point.

Shift Analysis

For a given laboratory test, subjects are usually classified into categories such as below the normal range, within the normal range, and above the normal range at pre- and post-treatment evaluation. Let n_{ij} be the number of subjects allocated to category i before treatment but to category j after treatment. Then the allocation of the subjects to the categories pre- and post-treatment can be set out in a 3×3 table (see Table 13.5.2). In Table 13.5.2 the entries on the main diagonal, namely n_{ij} for $i=j$, give the numbers of subjects for whom pre-treatment and post-treatment are in agreement. The off-diagonal cells show the disagreements. For example, n_{12} is the number of subjects whose laboratory values are below the normal range before treatment but are within the normal range after treatment, while n_{21} is the number of subjects whose laboratory values are within the normal range before treatment but fall below the normal range after the treatment.

A typical approach for the analysis of laboratory data as presented in Table 13.5.2 is to perform a shift analysis between pre- and post-treatment. The shift analysis for safety assessment is to test whether the number of subjects about which before and after treatment disagreed was similarly distributed across categories. The first analysis of interest is then to compare the distributions of row and column totals. In other words, it is of interest to examine whether the overall distribution of subjects to categories, irrespective of which subjects are allocated to which categories, is the same between pre- and post-treatment. For this purpose consider

$$d_i = n_{i\cdot} - n_{\cdot i},$$

Table 13.5.2 Allocation of Subjects to Categories of Pre-Treatment and Post-Treatment

		Post-treatment			
		Below	Within	Above	Total
Pre-treatment	Below	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
	Within	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
	Above	n_{31}	n_{32}	n_{33}	$n_{3\cdot}$
	Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n

Note: Below, within, and above mean below, within, and above normal range: $n = \sum_{j=1}^3 \sum_{i=1}^3 n_{ij}$.

where $i = 1, \dots, K$, the number of categories at pre- and post-treatment and

$$n_i = \sum_{j=1}^K n_{ij}, \quad n_j = \sum_{i=1}^K n_{ij}.$$

It can be easily verified that

$$\widehat{\text{var}}(d_i) = n_i + n_j - 2n_{ij}$$

for $i = j$ and

$$\widehat{\text{cov}}(d_i, d_j) = -(n_{ij} + n_{ji})$$

for $i \neq j$. Since the covariance matrix for the d_1, d_2, \dots, d_K is singular, we may consider omitting one of the d 's; that is let

$$\mathbf{d} = (d_1, \dots, d_{K-1})'.$$

Then, under the null hypothesis that the distributions of row and column totals are the same, statistic

$$\mathbf{d}' \sum^{\wedge-1} \mathbf{d},$$

is asymptotically distributed as a chi-square random variable with $K-1$ degrees of freedom, where $\hat{\Sigma}$ is an estimate of the covariance matrix of \mathbf{d} . Therefore we reject the null hypothesis that the distributions of row and column totals are the same at the α th level of significance if

$$\mathbf{d}' \sum^{\wedge-1} \mathbf{d} > \chi^2(\alpha, K-1) \quad (13.5.4)$$

where $\chi^2(\alpha, K-1)$ is the upper α th quantile of a chi-square distribution with $K-1$ degrees of freedom. Note that the rejection of the null hypothesis would imply that for at least one of the categories of the pre- and post-treatment differs in the proportions of the total sample allocated to that category. However, failure to reject the null hypothesis would not imply that the proportions of the subjects allocated to the different categories at pre- and post-treatment are necessarily the correct proportions, much less would it imply that the subjects are correctly classified.

In practice, it is also of interest to investigate whether the number of subjects at which pre- and post-treatment disagreed was distributed by them in a similar manner among other categories. In other words, it is of interest to compare frequencies in corresponding cells about the main diagonal in Table 13.5.2. Thus for $i < j$ the hypotheses of interest can be expressed as

$$\begin{aligned} H_0: \mu_{ij} &= \mu_{ji}, \\ \text{vs. } H_a: \mu_{ij} &\neq \mu_{ji}, \end{aligned} \quad (13.5.5)$$

where

$$\mu_{ij} = E(n_{ij}).$$

Under the null hypothesis of (13.5.5), the test statistic

$$X_{SM} = \sum_{i < j} \frac{(n_{ij} - n_{ji})^2}{(n_{ij} + n_{ji})} \quad (13.5.6)$$

follows a chi-square distribution with $\frac{1}{2}K(K - 1)$ degrees of freedom. Therefore, at the α th level of significance, we can reject the null hypothesis of (13.5.5) if

$$X_{SM} > \chi^2[\alpha, \frac{1}{2}K(K - 1)],$$

where $\chi^2[\alpha, \frac{1}{2}K(K - 1)]$ is the upper α th quantile of a chi-square distribution with $\frac{1}{2}K(K - 1)$ degrees of freedom. The test statistic given in (13.5.6) is known as the Stuart-Maxwell test.

Note that when $K = 3$, the Stuart-Maxwell statistic can be rewritten as

$$X_{SM} = \frac{\bar{n}_{23}d_1^2 + \bar{n}_{13}d_2^2 + \bar{n}_{12}d_3^2}{2(\bar{n}_{12}\bar{n}_{13} + \bar{n}_{12}\bar{n}_{23} + \bar{n}_{13}\bar{n}_{23})}, \quad (13.5.7)$$

where

$$\bar{n}_{ij} = \frac{n_{ij} + n_{ji}}{2}.$$

Example 13.5.1 To illustrate the statistical methods for comparing the distributions of row and column totals and frequencies in corresponding cells about the main diagonal as described above, consider the laboratory test results of a given parameter for a sample of 115 subjects in a clinical trial. The results were summarized in Table 13.5.3. From Table 13.5.3 we have

$$d_1 = 12 - 25 = -13$$

$$d_2 = 49 - 47 = 2,$$

$$d_3 = 44 - 43 = 11;$$

$$\widehat{\text{var}}(d_1) = 12 + 25 - 2(10) = 17,$$

$$\widehat{\text{var}}(d_2) = 49 + 47 - 2(32) = 32,$$

$$\widehat{\text{var}}(d_3) = 54 + 43 - 2(41) = 15;$$

Table 13.5.3 Allocation of Subjects for Hemoglobin Laboratory Test

		Post-treatment		
		Below	Within	Above
Pre-treatment	Total			
	Below	10	2	0
Within	15	32	2	49
Above	0	13	41	54
Total	25	47	43	115

Note: Normal ranges for hemoglobin test are 12.9–17.9 g/dL for male and 11.0–15.6 g/dL for female.

and

$$\begin{aligned}\widehat{\text{cov}}(d_1, d_2) &= -(15 + 2) = -17, \\ \widehat{\text{cov}}(d_1, d_3) &= -(0 + 0) = 0, \\ \widehat{\text{cov}}(d_2, d_3) &= -(13 + 2) = -15.\end{aligned}$$

Thus the matrix of variance and covariance of the d 's is given by

$$\begin{bmatrix} 17 & -17 & 0 \\ -17 & 32 & -15 \\ 0 & -15 & 15 \end{bmatrix}.$$

If we delete the last row and the last column of the matrix, we then have

$$\hat{\Sigma} = \begin{bmatrix} 17 & -17 \\ -17 & 32 \end{bmatrix}.$$

As a result the inverse matrix is given by

$$\hat{\Sigma}^{-1} = \frac{1}{255} \begin{bmatrix} 32 & 17 \\ 17 & 17 \end{bmatrix} = \begin{bmatrix} 0.125 & 0.067 \\ 0.067 & 0.067 \end{bmatrix}.$$

Thus

$$\mathbf{d}' \hat{\Sigma}^{-1} \mathbf{d} = 17.909 > \chi^2(0.05, 2) = 7.378.$$

Therefore, at the 5% level of significance, we can conclude that there is a significant difference in the distributions of row and column totals.

To compare frequencies in corresponding cells about the main diagonal, we apply the test statistic given in (13.5.6). The test result and the contributions of each cell to the overall chi-square test from the three pairs of cells are summarized in Table 13.5.4. The test result indicates that there is a shift in classifications between pre- and post-treatment. Table 13.5.4 also indicates that the greatest discrepancy occurs between cells n_{12} and n_{21} and between cells n_{23} and n_{32} . Therefore an attempt to improve agreement could be directed to eliminate the discrepancy.

As described in Chapter 9, another approach is to collapse Table 13.5.2 into a 2×2 table as illustrated in Table 13.5.5. Under Table 13.5.5 we can then apply the following

Table 13.5.4 Individual Contributions to Overall Chi-Square Test

Cells	Contribution to χ^2
n_{12} and n_{21}	9.941
n_{13} and n_{31}	0.000
n_{23} and n_{32}	8.067
Total	$\chi^2 = 18.008$

Table 13.5.5 Summary Table for Pre- and Post-Study Laboratory Test Results

		Post-study		
		Normal	Abnormal	Total
Pre-study	Normal	<i>a</i>	<i>b</i>	<i>a + b</i>
	Abnormal	<i>c</i>	<i>d</i>	<i>c + d</i>
	Total	<i>a + c</i>	<i>b + d</i>	<i>n</i>

McNemar's statistic to test the difference in changes:

$$X_M = \frac{\left\{ |p_2 - p_1| - 1/n \right\}^2}{\text{s.e.}(p_2 - p_1)} \\ = \frac{(|b - c| - 1)^2}{b + c}.$$

Note that

$$\begin{aligned} a &= n_{22}, \\ b &= n_{21} + n_{23}, \\ c &= n_{12} + n_{32}, \\ d &= n_{11} + n_{13} + n_{31} + n_{33}. \end{aligned}$$

If it is not significant, we can conclude that the change is due to chance. In other words, the treatment does not have any influence on the change.

Other Analyses

In addition to shift analysis, the ICH guidelines also suggest that the number or fraction of patients who had a parameter change of a predetermined size at selected time intervals be summarized in tables. For example, for BUN, it might be decided that a change of more than a predetermined value of BUN should be noted. For this parameter the number of patients having a similar or greater change would be shown for one or more visits, usually grouping patients separately depending on the baseline BUN. The advantage of this table is that changes of a certain size are noted even if the final value is not abnormal.

Another commonly used approach is the scatter plot with a 45° line. This approach is a graph combining the initial value and the on-treatment values of a laboratory measurement for each patient by locating the point defined by the initial value on the ordinate. If no changes occur, the point representing each patient will be located on the 45° line. A general shift to higher values will show a clustering of points above the 45° line. The scatter plot with a 45° line not only shows the baseline and most extreme on-treatment values but also helps identify potential outliers. Since the scatter plot with a 45° line usually shows a single time point for a single treatment, a time series approach for the analysis and interpretation of these plots is employed in comparing adverse events between treatment groups.

13.6 DISCUSSION

Unlike hypotheses for efficacy assessment, safety hypotheses usually cannot be specified a priori due to the following reasons. First, clinical trials are not designed to detect differences in safety outcomes. Second, clinical trials are usually not prepared to test hypotheses regarding rare safety events. However, as indicated by Levine (1996), failure to achieve statistical significance does not mean that a safety finding can be ignored.

In clinical trials, as was shown in this chapter, an adequate protocol must clearly define the responsibilities of the investigator in reporting and documenting all adverse events. The person(s) responsible for notifying appropriate authorities (regulatory, hospital ethics committee, etc) must be identified. When clinical and laboratory tests are used to monitor safety, parameters to be measured, timing and frequency, normal values for laboratory parameters, and definition of test abnormalities should be provided. For safety assessment, it is suggested that all patients entered into treatment who receive just one dose of the treatment should be included in the safety analysis. If a patient is not included, an explanation must be provided.

For measurement of risk of various adverse events type and data, Tremmel (1996) suggests that the crude rate or cumulative rate be used for short-term clinical trials and all types of adverse events. For long-term clinical trials, the hazard function or median survival time can be used as a meaningful measure of risk for absorbing adverse events. For recurring events with short duration, hazard functions or expected counts can be employed to have meaningful measures of risk. For recurring events with long duration, however, it is suggested that alternative approaches such as prevalence functions or expected proportions of time affected be used. For analysis and interpretation of adverse events, the FDA guidelines suggest that attention be given to (1) frequency of treatment-emergent events; (2) relevant Body System categories; (3) severity categories (if used); (4) relationship/causality (if used); (5) original terms used by the investigator (individual study report) and group related reactions, such as defined in one of the dictionaries (integrated safety summary). If a dictionary is not used in the study report, any synonymous reactions should be grouped together. If the study size permits, more common adverse events that seem to be drug related should be examined for their relationship to the levels of drug exposure and to the baseline characteristics. Laboratory findings can reflect an adverse event (ECG abnormality suggesting infarction, serious arrhythmia, etc).

For safety assessments in clinical trials, it is helpful to examine the correlation between the drug concentration data such as concentration at the time of an event, maximum plasma concentration, or area under the curve (if available) and adverse events or changes in laboratory variables.

Note that for the safety assessment of laboratory data, since a majority of the laboratory measurements are surrogates and reference ranges can vary from method to method, many different types of errors can occur during laboratory testing. The sources of errors can be the technician, an instrument, the environment, or reagents. As a result multiple laboratory methods within a statistical analysis should be avoided. In practice, it is strongly recommended that standard procedures for good laboratory practice (GLP) such as quality assurance/control, well-defined procedures, well-trained staff, and well-defined data management system be employed to minimize these potential errors.

In summary, there are still problems encountered in defining, capturing and evaluating safety-related data including adverse events and laboratory data in clinical trials. These

problems may be best resolved by the ICH in a standardized approach to defining, evaluating, recording, and summarizing safety-related data for a complete safety assessment of the drug product under investigation. It should be noted that both FDA and ICH Guidelines pointed out that it is not intended that every adverse event be subjected to rigorous statistical evaluation. As a result, the analysis of adverse events is basically descriptive in nature.

14

PREPARATION AND IMPLEMENTATION OF A CLINICAL PROTOCOL

14.1 INTRODUCTION

As mentioned in Section 1.4, a clinical trial protocol is the document that specifies the research plan for a clinical investigation. A protocol is the most important document to ensure quality control of a clinical trial. It is also a document for communications among centers and research staff, such as clinical investigators, study nurses, pharmacists, and all others who are involved in a clinical trial. A clinical trial protocol is also required to be reviewed both internally and externally. Internal review may involve the institutional review board (IRB) or ethics committee (EC) within the medical institution from which the protocol is originated. In most countries, a clinical trial can be conducted only after the health authorities or national IRB approves its protocol. In addition, a protocol is a legal document that specifies the legal responsibilities of all parties participating in a clinical trial. Finally, although a clinical trial is a research plan for a scientific investigation, its experimental units are humans and, hence, the ethical obligations of all personnel involved in a clinical trial should be clearly stated in the protocol. In summary, a clinical trial protocol is a plan that meets scientific and ethical requirements as well as good clinical practice (GCP) (see, e.g., ICH E6).

The ultimate goal of a clinical trial is to prospectively design an investigation to make an unbiased inference with possibly the best precision to scientifically answer clinical questions with respect to a target patient population. As pointed out in Chapter 1, the most important elements in a clinical trial are experiment unit, treatment, and evaluation. Therefore, a clinical trial protocol should provide a plan with complete specifications for the following:

- research question
- target patient population

Design and Analysis of Clinical Trials: Concepts and Methodologies, Second Edition

By Shein-Chung Chow and Jen-pei Liu

ISBN 0-471-24985-8 Copyright © 2004 John Wiley & Sons, Inc.

- design characteristics
- treatment characteristics
- data collection and analysis
- ethical obligations
- legal responsibilities
- research management

On the other hand, a well-written protocol should be scientifically valid, ethical, flexible, structured, logical, and complete. From the above discussion, it is not easy to develop a good clinical trial protocol. However, the most difficult parts for development of a clinical trial protocol include:

1. Formulation of and development of a set of important and yet feasible scientific/medical questions
2. Necessary resources (patients, funds, time, personnel, equipment) to answer the questions

Although a clinical trial protocol is well prepared, thoroughly thought out, and adequately written, its implementation can still face lots of difficulties and obstacles due to some unforeseen situations and unanticipated issues during the conduct of the clinical trial. Some of these issues are relatively minor, which are the so-called protocol *deviations*, and some of them are quite major, which are usually referred to as *protocol violations*. Some protocol violations are very serious, which may lead to *misconduct* or *fraud* in clinical trials. Therefore, monitoring and audit during the trial by the sponsor are required to ensure that a clinical trial is conducted according to the clinical trial protocol and GCP. An official inspection by the health authorities is usually conducted to evaluate quality and performance of a clinical trial after the completion of the trial. All of these efforts are aimed to ensure that trials are conducted and data are generated, documented, and reported in compliance with the trial protocol, GCP, and applicable regulations set forth by the health authorities. The purpose is to make sure that the subjects enrolled in trials are adequately protected and that valid data are generated to answer the scientific questions stated in the clinical trial protocol.

In the next section, the structure and components of a clinical trial protocol will be introduced. Points to be considered and some common pitfalls arising during development and preparation of a clinical trial protocol will be discussed in Section 14.3. Commonly occurring departures from the protocol during the conduct of a clinical trial will be given in Section 14.4. Section 14.5 briefly discusses the roles of monitoring, audit, and inspection for quality assurance of clinical trials. Quality assessment of a clinical trial will be provided in Section 14.6. Discussion and final remarks are given in Section 14.7.

14.2 STRUCTURE AND COMPONENTS OF A PROTOCOL

As indicated in Section 1.4, the minimum requirements for the protocol of a clinical trial are provided in Section 312.23 of 21 CFR. In addition, Table 1.4.2 provides an example for format and contents of a protocol for a well-controlled clinical trial. A protocol cover sheet suggested by the FDA provides a tabular summary of information and characteristics for the trial. On the other hand, a protocol synopsis with a length of one to two pages can depict a narrative summarized description of a complicated and lengthy protocol. The table

of contents of the protocol not only provides the structure and framework of the protocol, but also it serves as an index for each section of the protocol. Basically, a protocol can be structured into three major components: the scientific, ethical, and administrative parts. According to the ICH E6 guideline on *Good Clinical Practice: Consolidated Guidance* (ICH, 1996), the administrative parts consist of general information, organization of research teams, communication schemes among participants involved in the trials, shipping plan for specimen, data transfer network, obligations of clinical investigations, medical institutions, and sponsors, investigational drug accountability, case report forms, study registry, record retention, financing and insurance, publication policy, signature of investigators, confidential statements, and other administrative matters.

The research team of the clinical trial usually consists of the team of clinical staff at the study site, which includes clinical investigators, residents, research fellows, research nurse, pharmacists, laboratory technicians, and clinical assistants. The leader of this team is referred to as the *principal investigator* (PI). Individuals such as residents, research associates, and research fellows who are supervised by the principal investigator to perform critical trial related procedures and to make important trial related decisions are sometimes referred to as *subinvestigators*. If a clinical trial is sponsored by a pharmaceutical company, another research team is formed by the staff of the pharmaceutical company that may include project clinicians, a statistician, a pharmacokineticist, and a project manager. For a large-scale multicenter trial, a steering committee is usually formed to set up the research strategy, to make key decisions, and to oversee the activities of the trial. One of the principal investigators is usually selected as the chairperson of the steering committee. The steering committee is critical for successful management of large, long-term, and complicated clinical trials such as the PLCO trial discussed in Chapter 7. In practice, it is not uncommon to have a data and safety monitoring committee (DSMC) as another research team for clinical trials. As a matter of fact, the NIH requires that all clinical trials sponsored by the NIH have a DSMC. Furthermore, most phase 2 and phase 3 trials sponsored by the pharmaceutical industry also include DSMC. One key and crucial distinction between and DSMC and other research teams is that all members of the DSMC should not be involved with the conduct of the trials. In addition, the DSMC should function independently.

According to the ICH E6 GCP guideline, the general information provided by the protocol should include the following:

- Protocol title, protocol identifying number, and date. Any amendment(s) should also bear the amendment number(s) and date(s)
- Name and address of the sponsor and monitor (if other than the sponsor)
- Name and title of the person(s) authorized to sign the protocol and the protocol's amendment(s) for the sponsor
- Name, title, address, and telephone number(s) of the sponsor's medical expert (or dentist when appropriate) for the trial
- Name and title of the investigator(s) who is (are) responsible for conducting the trial, and the address and telephone number(s) of the trial site(s)
- Name, title, address, and telephone number(s) of the qualified physician (or dentist, if applicable) who is responsible for all trial site-related medical (or dental) decisions (if other than investigator)
- Name and address(es) of the clinical laboratory(ies) and other medical and/or technical department(s) and/or institutions involved in the trial.

The scientific part contains the specific applications of the most methodological tools covered in the previous chapters of this book. They include background information, the objectives of the trials, definition of target patient population according to inclusion and exclusion criteria, the trial design, intervention or treatment under investigation for the trials, methods for evaluation of efficacy and safety, reporting procedures for serious and unexpected adverse events, and sample size determination, data collection, and a statistical plan for analyses. If the treatment under investigation is a pharmaceutical product, then the ICH E6 GCP guideline indicates that the following background information should be provided:

- Name and description of the product(s)
- A summary of findings from nonclinical studies that potentially have clinical significance and from clinical trials that are relevant to the trial
- Summary of the known and potential risks and benefits, if any, to human subjects
- Description of and justification for the route of administration, dosage, dosage regimen, and treatment period(s) for the treatment evaluated in the trial
- A statement that the trial will be conducted in compliance with the protocol, GCP, and applicable regulatory requirement(s)
- A brief description of the population to be studied and its rationales
- References to literature and data that are relevant to the trial, and that provide background for the trial

The most important and yet difficult part during the development of a clinical trial protocol is the objective section that lays out the goals for the trial. From Example 3.2.1 to 3.2.7 of Section 3.2, the objectives of clinical trials are formulated through the experimental units, treatments, and evaluation in conjunction with the mechanism for data generation and the intentions of the trial. In the primary objective of Example 3.2.7, the experimental units are patients with atrial fibrillation. The treatments include the investigational product and the active comparator. In general, the dose and route of administration should be also given in the objective. The evaluation endpoint is the all-stroke (fatal and nonfatal) and systemic embolic events. The mechanism for data generation is the design used in the trial, which is a randomized parallel group design. Finally, the intention of the trial is to show that the investigational drug is not inferior to that of the active comparator for the prevention of all-stroke (fatal and nonfatal) and systemic embolic events. Therefore, this trial is a secondary prevention and noninferiority trial. Sometimes, it is useful to state the treatment duration and the length of follow-up period. It would be more complete if the expected event rate and noninferiority margin can be stated in the primary objective. However, the objective would become too long and yet confusing with too much information. Therefore, it is very difficult to reach a proper balance among comprehensiveness, completeness, clearness, and conciseness. A sound statement regarding the objectives of a clinical trial protocol, however, should at least include experimental units, treatment, evaluation, design, and intention.

Depending on the relative importance or priority of study endpoints, as shown in Example 3.2.5 to 3.2.7, the objectives of a clinical trial can also be classified into the primary, secondary, and tertiary objectives. In general, there should be only one primary objective and several secondary and tertiary objectives. The endpoint stated in the primary objective is the primary endpoint for the trial that is used for sample size calculation. In practice, there may be more than one primary objective, each with different primary endpoints. In such a case, the multiplicity issue discussed in Chapter 12 should be addressed in appropriate sections (e.g., the

objective section and the statistical section) of the clinical protocol. In addition to inclusion and exclusion criteria discussed in Section 3.3 for the target patient population, the ICH E6 GCP guideline also provides the following subject withdrawal criteria:

- When and how to withdraw subjects from the trial/investigational product treatment
- The type and timing of the data to be collected for withdrawn subjects
- Whether and how subjects are to be replaced
- The follow-up for subjects withdrawn from trial treatment

As pointed out in the ICH E6 GCP guideline, the integrity of the trial and the credibility of the data obtained from the trial depend considerably on the design employed in the trial. Therefore, the ICH E6 GCP guideline suggests that the design section should include the following:

- A description of the type/design of trial to be conducted (e.g., double-blind, placebo-controlled, parallel design) and a schematic diagram of trial design, procedures, and stage.
- A description of the measures taken to minimize/avoid bias, including randomization and blinding.
- A description of the trial treatment(s) and dosage and dosage regimen of the investigational product(s). Also includes a description of the dosage form, packaging, and labeling of the investigational product(s).
- The expected duration of subject participation, and a description of the sequence and duration of all trial periods, including follow-up, if any.
- A description of the *stopping rules* or *discontinuation criteria* for individual subjects, parts of trial, and entire trial.
- Accountability procedures for the investigational product(s), including the placebo(s) and comparator(s), if any.
- Maintenance of trial treatment randomization codes and procedures for breaking codes.
- The identification of any data to be recorded directly on the CRF's (i.e., no prior written or electronic record of data), and to be considered to be source data.

The treatment section should provide all relevant information of all treatments administered to the subjects during the trial. These include the name(s) of all the product(s), the dose(s), the dosing schedule(s), route mode(s) of administration, treatment period(s), and follow-up period(s) for subjects for each investigational product treatment/trial treatment group/arm of the trial. In particular, quantity, frequency, time, and duration should be specified for the investigational products that are orally administered. For the IV investigational products, the preparation, reconstitution procedures, the infusion rate, loading, and maintenance doses should be also described clearly in this section. For the treatments that require titration, dose, schedule, criteria, and its based endpoints should be specified in this section. For the trials evaluating chemotherapy in cancer patients, schema for dose modifications due to adverse events or toxicities should also be described in detail in the treatment section. On the other hand, the treatment section should also include the procedures for monitoring subject compliance and methods for improving compliance. In order to avoid bias in estimation of treatment effects, the treatment section should describe the procedures to collect the

Table 14.2.1 Desirable Characteristics of Efficacy Parameters

-
1. Easy to administer
 2. Rapid to administer
 3. Little or no training necessary to administer
 4. Easy to interpret
 5. Rapid to interpret
 6. Little or no training necessary to interpret
 7. Sensitive to changes elicited by treatments under investigation
 8. Insensitive to efforts brought about by treatments that might interfere with the efficacy parameter
 9. Low rate of false-positive responses
 10. Low rate of false-negative responses
 11. May be used multiple times without a training effort
 12. Results are reproducible
 13. Results are valid
 14. Interpretation is correlated with other efficacy parameters
-

Reproduced from Spilker (1991).

information on concomitant medications. In addition, medication/treatments permitted (including rescue medications) and not permitted before and/or during the trial should be specified in detail with prespecified time intervals.

Basically, the section of assessment of efficacy can be divided into two parts. The first part includes definitions and specifications of efficacy parameters. In this part, depending on clinical relevance and importance, efficacy parameters can be classified further in the primary, secondary, and tertiary efficacy parameters in accordance with the primary, secondary, and tertiary hypotheses specified in the objective section. Spilker (1991) gave a list of desirable features of efficacy parameters, which is reproduced with permission in Table 14.2.1. For each efficacy parameter, a precise definition should be clearly given in detail. In particular, the definitions of some derived efficacy parameters such as response rate should include the efficacy parameters on which they are based and their calculations. In addition, if some efficacy parameters are derived from some criteria or instrument scale, they should be validated for their validity, reliability, and reproducibility. Some examples include RECIST for evaluation of tumor response in cancer chemotherapy (Therasse et al., 2000), Hamilton depression scale for evaluation of treatment in patients with depression, NIHSS for evaluation of treatment in patients with stroke (NINDS, 1995), and ADAS for evaluation of treatment in patients with probable Alzheimer's disease (Folstein et al., 1975).

The second part should provide a detailed narrative description for methods and timing for assessing, recording, and analyzing efficacy parameters. In addition to selection of reliable, valid, and reproducible efficacy parameters, the time points for measurement of efficacy parameters should be selected to demonstrate efficacy in relation to the time of administration of treatment under investigation described in the treatment section, e.g., just prior to dosing. In addition, it is crucial to define the times for evaluating inclusion/exclusion criteria that are based on efficacy parameters and for establishing the baseline values for efficacy parameters. Appendices should be provided for detailed narrative description of procedures if evaluations and measurements for some efficacy parameters require special equipments, training, or techniques. Some examples are bone densitometry for evaluation of bone mineral density and NIHSS, which requires additional training for neurologists in assessment of improvement by the treatment for the patients with stroke.

Similar to the section for assessment of efficacy, the section for assessment of safety can also be divided into two parts. The first part provides specifications of safety parameters that include adverse events, laboratory evaluations, vital signs, and some specific safety parameters. The second part includes the methods and timing for assessing, recording, and analyzing safety parameters. The definition of adverse events should be given in details. The information collected for adverse events should also be specified clearly in this section. This includes description of events, onset date and time in relation to the treatment, nature of the adverse events, its duration, intensity, seriousness, relationship to the treatment, actions taken, and resolution of adverse events. In addition, to avoid a possible overestimation of the incidence rate of an adverse event, the information of adverse events should be obtained by indirectly questioning using a nonleading question. One should not, in the protocol, prepare a list of all possible adverse events and ask patients whether they experience them. The coding system for adverse events described in Chapter 13 to be used in the trial should be also prespecified in this section. On the other hand, the time interval for collection of adverse events even after the completion of the trial treatment should be specified in this section.

Serious adverse events (SAE) and unexpected adverse events are the most important safety information to be collected and evaluated for a clinical trial that also require expedited reporting to the medical institute's IRB and health regulatory authority and the industry sponsors. Therefore, a detailed and precise narrative description of the definition and procedures for obtaining and reporting SAE should be provided in this section. Currently, the ICH definition of serious or unexpected adverse events is used for most of clinical trials sponsored by the pharmaceutical companies. Currently, the time frame for expedited reports of serious or unexpected adverse events suggested by the ICH is also adopted by most of the trials sponsored by the pharmaceutical industry. In addition, there should be a statement indicating that the principal medical monitor or sponsor should be notified promptly by phone of any death or serious or unexpected adverse event, even though they are not product-related. The name, address, business phone number, business hour, fax number, and email address of the principal monitor should be given in the protocol along with the same information for the physician on duty for nights, weekends, and holidays.

Other safety parameters include laboratory evaluations, vital signs, and some special safety parameters. In general, routine laboratory evaluations include hematology, blood chemistry, and urinalysis. The time and procedures for collecting, labeling, and shipping the blood and urine samples should be described in this section. For multicenter trials, whether a central laboratory is used for analyzing the blood or urine samples should be specified. If samples are to be analyzed at site laboratories, a standardization procedure among different laboratories should also be depicted in this section. Vital signs usually include systolic and diastolic blood pressures, heart and respiratory rates, and sometimes body temperature. Examples for special safety evaluations include ECG and pap smears tests. Again, the protocol should specify the procedures and time points for evaluation of these safety parameters.

However, there should be another section on other clinical evaluations that include medical history, medication history, and physical examination. A section on the conduct of the study can be very helpful. In this section, narrative descriptions for every evaluation of all efficacy and safety parameters performed at each visit for all visits are provided by visit. In addition, a flowchart is given in appendices to provide a summary of all evaluations by time points. In this flowchart, columns represent the visits or time points; the rows represent the evaluations of efficacy and safety parameters. Therefore, each entry in the flowchart indicates a particular efficacy or safety evaluation to be performed at a specific time point.

The section on statistics should include the following parts: statistical hypotheses and sample size determination, definition of analysis set, analysis of demographic data and baseline characteristics, analysis of efficacy parameters, and analysis of safety parameters. First, the objectives in the beginning of the protocol should be translated into statistical hypotheses. Based on the expected difference on primary efficacy parameter between treatment arms and their corresponding anticipating variability, the sample size for the trial can be determined by the corresponding statistical method used for analysis of the primary efficacy parameter to achieve a predetermined level of power at a prespecified significance level. Clinical relevance for selection the difference between treatment arms or equivalence/noninferiority margin for determination of sample size should be justified. The sample size calculation should also take dropout rate into consideration. If the trial is a multicenter trial, the number of subjects projected for each site should be specified. In addition, the projected accrual rate and anticipated completion of enrollment should be provided.

The statistics section should clearly define the analysis sets from which the conclusions of the trial are derived. The definition of analysis sets and the criteria for selection of subjects into analysis sets should be clearly described and specified in this section. The methods for analysis of each efficacy and safety parameters should be provided with relevant references if necessary. If interim analyses are planned for the trial, the timing of interim analyses and the methods for determination of spending significance level at various time points for interim analyses and the corresponding stopping boundaries should also be provided in detail. The procedures for early termination of the trial based on planned interim analyses should also be prespecified in the protocol. Furthermore, if necessary, the methods for sample size reestimation should be also given. However, the procedures to avoid or eliminate bias due to the planned or unplanned interim analyses and sample size reestimation should be specified in the protocol to protect the integrity of the trial. The statistics section should provide the procedures for addressing the issues of multicenter trial, use of covariates (both subject-specific covariates and time-dependent covariates), missing values, and multiplicity.

The last component of the protocol is the ethical section. Because the experimental units in clinical trials are human subjects, the ethical section is probably the most important component in the protocol for protection of the trial subjects. This component should include the detailed review procedures both internally and externally. In addition, a copy of the proposed informed consent form should be included as one of the appendices. Other procedures for elimination of any unnecessary risk from the trial subjects should be provided in this section.

14.3 POINTS TO BE CONSIDERED AND COMMON PITFALLS DURING DEVELOPMENT AND PREPARATION OF A PROTOCOL

Development and preparation of a clinical trial protocol begin with scientific/medical questions and end with a comprehensive and complete investigational plan that is doable and feasible to test and verify clinical hypothesis. A good and well-written protocol can not only achieve its goal by answering the scientific/medical questions if it is executed according to the plan, but also prepare different contingency plans to handle and resolve various unforeseen or unexpected situations and issues during the implementation of the trial protocol. Therefore, it is a challenging and complicated process that requires careful and thoughtful planning. The Women's Health Initiative Clinical Trial (1998) and the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial (2000) provide two excellent examples of

Table 14.3.1 Points to Consider in Development and Preparation of a Clinical Protocol

-
1. Assembling a project team of clinical experts
 2. Selection of a management approach for development and preparation of a clinical protocol
 3. Development of the objectives for the protocol
 4. Definition of the target patient population
 5. Selection of the proper design for the trial
 6. Selection of efficacy and safety parameters
 7. Specification of the time for randomization and for baseline evaluation
 8. Collection of information on adverse events
 9. Ethical considerations and issues, including informed consent form and IRB/EC review processes
 10. Regulatory, patent, legal considerations
 11. Preparation, packing, labeling, and dispensing of investigational products
 12. Preparation of case report forms and instructions for trial subjects, investigators, and other relevant trial personnel
 13. Inconsistency within the protocol
 14. Inadequate language and writing style for the protocol
-

complexity in development and preparation of clinical trial protocols. Table 14.3.1 provides a list of key issues that must be considered for development and preparation of a clinical trial protocol.

As it can be seen from Table 1.4.2, a protocol is a product of a joint effort by experienced experts from different disciplines. Therefore, the starting point for development of a clinical protocol is to assemble a project team of qualified professionals from different areas. They may include clinical physicians specialized in the therapeutic area under study; medicinal chemists who are familiar with the properties of the investigational products; pharmacokinetics and pharmacodynamic (PK/PD) experts who understand the metabolism and mechanism of the products; toxicologists who are familiar with animal safety profiles; pharmacists who are experts in preparing, packaging, and labeling of the investigational products; biostatisticians who have the knowledge of choosing a valid design and appropriate statistical methods for sample size determination and analyses; and database experts who are experienced in management of complex clinical data. The next step is to select a management approach. In general, preparation and implementation require expertise from different disciplines. Decision by consensus seems a reasonable approach. However, the most important step in forming a project team is to select a project leader who respects and appreciates different disciplines and understands the complexity of their individual roles involved in the clinical trials. The ability and efficiency of communication is also another key element in successful management of a project team for development and preparation of a clinical trial protocol.

Once the project team is formed, the first task is to develop the study objectives of the protocol. The different study objectives of a clinical trial with examples have been discussed in Chapter 3. The following lists mistakes that are commonly made when developing the study objectives by a project team:

- The trial objectives are too general, too vague, or too ambiguous such as Example 3.2.1 to 3.2.3
- The trial objectives are too ambitious to achieve, e.g., a reduction of incidence rate 1 out of 100,000 for a certain event by 50%

- The trial objectives either require a complicated statistical design or contain too many parts or phases that are difficult to implement or demand huge resources to complete

As mentioned earlier in Chapter 3, defining the target patient population requires an adequate balance between variability and recruitment. If inclusion and exclusion criteria are very stringent, the resulting target patient population may be very homogeneous with respect to some basic characteristics. However, due to very stringent eligibility criteria, accrual can be very slow. On the other hand, if inclusion and exclusion criteria are too loose, data obtained from the trial may be very heterogeneous and are of less clinical meaning.

After the target patient population is defined, the next task is to choose a proper design for the trial. The study design consists of statistical design, randomization, blindness, selection of controls, and duration of the trial. Table 14.3.2 provides a list of possible pitfalls for selection of a clinical trial design. The first pitfall is that no statistician is on the project team. In practice, it is crucial to have a biostatistician in the project team for selection of an adequate statistical design to appropriately address the study objectives. Another common pitfall during development and preparation of a clinical trial protocol is either too many or too few efficacy and safety parameters selected for evaluation of treatments. Too few parameters may result in insufficient clinical data to address the study objectives. On the other hand, a shotgun approach may generate too much fruitless data and provide inconclusive results. Another important issue to be considered during the development of a protocol is the time of randomization and evaluation of baseline characteristics of the patients. This issue is particularly important when a trial is designed with a run-in phase or an enrichment phase. For safety concerns, other common pitfalls are that excessive blood samples are planned to collect from subjects or the proposed fasting period is exceptionally long. In addition, because English is currently the most frequently spoken language in the world, the drug label and dispensing instructions are prepared only in English for the trials to be conducted in the countries or areas where patients or even some nurses cannot read or speak English proficiently.

Other common pitfalls include inadequate or ambiguous language with lots of adjectives used in the protocol and inconsistency within the protocol. For example, one inclusion criterion states that the patients should be between 18 and 65 years of age. However, this inclusion criterion is not clear and precise concerning whether patients of 18 years or

Table 14.3.2 Frequent Pitfalls in Selection of a Trial Design during the Development and Preparation of a Clinical Trial Protocol

-
1. No statistician in the project team to discuss and consult with in details on design issues
 2. The selected design fails to address the study objectives of the trial
 3. The selected statistical design is too complicated to implement
 4. The proposed design is inadequately powered
 5. The proposed lengths of screening, running, or treatment periods are either too long or too short for an appropriate conduct of a clinical trial to be implemented
 6. Placebo is not selected for the control when it should
 7. Blindness is not proposed when it should and is feasible
 8. Randomization procedure is not described in the protocol, and the statistician is not consulted with the issues of randomization
 9. Inadequate selection of dose regimen or doses for the trial
 10. The chosen margin of equivalence or inferiority trials is inappropriate to demonstrate the effectiveness of the treatment
-

65 years of age should be enrolled. A better statement may be that the age of the patients is between 18 and 65 inclusively. The following example illustrates inconsistency within the protocol. “The treatment is a **4-hour** intravenous infusion. Vital signs will be monitored every 15 minutes for the first 2-hours after initiation of infusion and every hour for the rest of the **12-hour** IV infusion.” Therefore, it is not clear from these two sentences whether the treatment is 4-hour IV infusion or 12-hour infusion. The minimal writing requirements are the language should be clear, concise, precise, and consistent without excessive adjectives or adverbs and long sentences.

14.4 COMMON DEPARTURES FOR IMPLEMENTATION OF A PROTOCOL

The U.S. FDA conducts approximately 1,050 inspections of clinical trials each year. Among these inspections, about 700 involve investigators, 250 IRBs, and 100 institutional facilities (Steinbrock, 2002a). Table 14.4.1 provides a list of the most common deficiencies found during the inspection. One of them is the protocol nonadherence, which occurred 26% of the time for the inspections conducted between 1977 and 1990, as reported by Lisook (1990). This situation did not improve at all for the trials conducted recently. Woollen (2000b) reported that the protocol nonadherence occurred 27% of the time for the trials inspected in 1999. Therefore, departures from the protocol during implementation and conduct of a clinical trial is a common problem that occurred in around 25% of the trials inspected during the last quarter century.

Depending on severity and intensity, departures from the trial protocol may be classified into three categories, namely, mild, moderate, and severe. *Deviation* from the protocol is the mildest form of departure from the protocol. For example, a small group of the patients fails to return for a visit or patients are not in full compliance of dosing schedule specified in the protocol. Another example is that a patient is scheduled to return for a 28-day visit as specified in a protocol for evaluation of a new drug in treatment of patients with hypertension. This protocol has specified a ± 4 -day window for each 28-day visit. However, due to a family emergency, this patient had this visit 5 days after the scheduled 28 days after the previous visit, i.e., one day after the allowable interval. Another deviation that occurs much less frequently during the conduct of the trial is missing a random code during assignment of treatment to patients. For a randomized trial, prepackaged treatments should be assigned to patients in a sequential order according to the random codes. For example, patient #10 is the tenth patient who satisfies the inclusion/exclusion criteria and should be given the

Table 14.4.1 Common Deficiencies Found during Inspection

Deficiencies	Year	
	77–90	99
Consent form	54%	8%
Protocol nonadherence	26%	27%
Inadequate or inaccurate records	21%	20%
Drug Accountability	25%	9%
AE		15%

Source: Lisook (1990) and Woolen (2000b).

treatment marked with number 10 on the label. However, if many patients are registered and enrolled at the same time, some mishaps can happen; instead of the treatment marked with number 10, patient #10 is assigned to the treatment marked with number 11. Even worse, if the study uses a stratified randomization, the treatment with number 10 from stratum A is assigned to patient #10 in stratum B. In general, few incidences of protocol deviations may have minimal impact on integrity, results, and conclusion of the trial. However, frequent occurrence of protocol deviations demonstrates lack of quality control for the protocol and study. In addition, the cumulative effect of frequent occurrence of protocol deviations will have a serious effect on a valid inference provided by the trial.

Violation of protocol is moderate departure from the protocol during implementation that might be of serious consequences in the integrity, quality, and validity of clinical trials. One example is nonadherence of the time of randomization. For example, a trial was conducted to investigate the effectiveness and safety of a new drug in treatment of allergic seasonal rhinitis. This study consisted of a screening visit, a 7-day run-in period, and a 4-week of treatment period. The primary endpoint was the change of the total symptom score at 4-week visit from the baseline. The protocol specified that baseline is the end of the 7-day run-in period, which is also the time for randomization and patients would be randomly assigned to treatment if the total symptom score is at least 12 at the time of randomization. To increase the enrollment, patients were actually randomized at the screening visit if their total symptom score was greater than or equal to 12. However, there was a 7-day run-in period between the screening and the baseline visit. The total symptom score evaluated at baseline for more than 60% of the patients fell below 12. In other words, more than 60% of the patients did not have the disease severity specified by the protocol at baseline and these patients do not represent the target patient population that the trial intended to study. Another example of protocol violation is noncompliance of prohibited concomitant medications. Suppose that a trial is conducted by the department of orthopedics in a large medical center to investigate the efficacy and safety of a new NSAID in treatment of pain in the patients with osteoarthritis. The protocol specifies that no other NSAID in any dosage form be allowed 7 days before the study and for the entire duration of the study. Due to lack of effective and efficient communications among different departments within the center, more than 40% of the patients in the trial were prescribed other NSAIDs when the patients sought medical help from other departments. As a result, an unbiased estimate of the effectiveness and safety of the new NSAID using the data generated by this trial is impossible. These two examples show the devastated impact of protocol violations on integrity and quality of the trial.

Most protocol deviations and violations result from issues and difficulties associated with the actual implementation of the protocol. The occurrence of protocol deviations and violations increases as the difficulty of implementing the protocol increases and the feasibility of the protocol decreases. Once these problems in conduct of the study are identified, they should be brought to the attention of the steering committee immediately to examine whether these issues are serious enough to warrant a protocol amendment. Protocol amendments can also be classified into two types. The first type is the administrative amendments, which include notification of changes of study personnel and modification of some administrative procedures. Administrative amendments do not have an impact on the scientific integrity and conduct of the trial. Therefore, in general, no official approval by the IRB or the health regulatory authorities is required. On the other hand, if amendments involve modification of scientific or ethical components of the protocol, then the proposed amendments must obtain the official approval from the IRB and/or health regulatory authorities before they can be implemented. Examples of these amendments include modification of inclusion/exclusion criteria

Table 14.4.2 Examples of Misconduct in Clinical Trials

Year	Case
1981	Darsee case (Relman, 1983; Culliton, 1983)
1994	Poisson (NSABP) case (Angell and Kassire, 1994; Broder, 1994; Bivens and Macfarlane, 1994)
1995	Bezwoda (high-dose chemotherapy for metastatic breast cancer) case (Grady, 2000; Weiss et al., 2001)
1996	University of Rochester (Steinbrock, 2002a), a healthy volunteer died
1999	University of Pennsylvania (Shalala, 2000), a patient died in a gene therapy trial
1999	Fiddes Case (Eichenwald and Kolata, 1999; Woollen, 2000b)
2000	Demiroglu (interferon in Behcet's disease) case (Horton, 2000; Horton, 2001).
2001	Case Western University (Steinbrock, 2002b), a control subject died
2001	Johns Hopkins University (Steinbrock, 2002a), a healthy volunteer died in an asthma trial

to increase the accrual of the study, sample size reestimation, modification or addition of primary endpoints, and revision of the informed consent form with updated information.

Protocol deviations or violations are unintentional departures from the protocol because of some mishaps or mistakes made during implementation of a clinical trial protocol. However, if departures from the protocol are intentional, then they are referred as to *misconduct or fraud* in clinical trials. Although the occurrence of misconduct and fraud in clinical trials is very rare, they did happen. Table 14.4.2 lists some infamous misconduct in clinical trials. Some of these misconducts involve fabrication or falsification of the data.

Example 14.4.1 Poisson (NSABP) Case

In May 1991, Dr. Roger Poisson, M.D., one of investigators for the National Surgical Adjuvant Breast and Bowel Project (NSABP), located at L'Hôpital Saint-Luc, University of Montreal, Quebec, Canada was found to falsify or fabricate the data. Dr. Poisson participated in 22 NSABP trials with a total enrollment of 1,500 patients. There were a total of 115 instances of data falsification or fabrication in eligibility criteria from 99 patients, such as falsified estrogen receptor value and alteration of dates of surgery and biopsy. Although this misconduct only included changes of data with respect to inclusion/exclusion criteria to increase trial accrual and no outcome data had been falsified or fabricated, Dr. Poisson was charged with scientific misconduct and was disqualified for life as a clinical investigator for conducting any clinical trials involving investigational drugs by the U.S. FDA.

Example 14.4.2 The Demiroglu Case

The other example of fraudulent and forged evidence is the paper by Demiroglu et al. (2000) regarding the efficacy of interferon for treatment of Behcet's disease published in *Lancet*. Immediate after the paper was published, several authors wrote separate letters to the Editor of *Lancet*, Robert Horton, M.D., indicating that they had not participated in the study and had not signed any copyright agreement. Dr. Horton then asked the Dean of Hacettepe University Medical School in Turkey to investigate this case. The investigation revealed that (1) the signatures on the copyright agreement had been forged, (2) no approval from the ethics committee was obtained, (3) patients in the alleged trials had not signed the informed consent form, and (4) some data had been fabricated and falsified. Eight months after the paper was published, *Lancet* (Horton, 2000) retracted the paper.

Example 14.4.3 Bezwoda Case

One of the most notorious examples of outright fraud in clinical trials is the case of Werner Bezwoda, M.D., a South African oncologist. The results of his 1995 paper on high dose chemotherapy for treatment of the patients with metastatic breast cancer published in the *Journal of Clinical Oncology* were so promising that the U.S. NCI planned large clinical trials based on his results. To assess the integrity of the trial and the quality of the data of the study, the U.S. NCI sent a team to South Africa to conduct an on-site audit for the Bezwoda's study. Despite the resistance and uncorporation from Dr. Bezwoda, the audit team concluded that patient records and diagnoses were unverifiable. Most patients were in fact ineligible, and verifiable data were not sufficient. Consequently, the paper was retracted 6 years after its publication in the *Journal of Clinical Oncology*.

Example 14.4.4 The Fiddes Case

Misconduct and fraud in clinical trials are not limited only to academia. Although the trials sponsored by the pharmaceutical industry for registration of new drugs are tightly regulated and monitored by both the industry and the U.S. FDA, misconduct and fraud still occur. The most infamous case is the fraud committed by Dr. Fiddes (Eichenwald and Kolata, 1999; Woollen, 2000b). Dr. Fiddes was the president of the Southern California Research Institute, a contract research organization located in Whitter, California. Since the early 1990s, the Southern California Research Institute has conducted over 200 clinical trials for as many as 47 pharmaceutical companies. Some examples of misconduct committed by Dr. Fiddes included making up fictitious study subjects, fabricating laboratory results by substituting clinical specimens, manipulating laboratory instrumentation, and prescribing prohibited medications to trial subjects for manipulation of data. The most notorious example of the fraud by Dr. Fiddes is a clinical trial investigating the efficacy/safety of an antibiotic. One of the inclusion criteria is the requirement that the patients have a certain type of bacteria growing in their ears. Dr. Fiddes bought the bacteria from a commercial supplier and shipped it to testing laboratories and said that they had come from his patient's ears. In August 1997, Dr. Fiddes pleaded guilty to a felony charge of conspiracy to make false statements to the U.S. FDA in connection with the drug approval process. He was sentenced to 15 months in federal prison and ordered to pay US\$800,000 in restitution. In addition, Dr. Fiddes was disqualified as a clinical investigator for life by the U.S. FDA.

The examples of misconduct or fraud are limited to fabrication, falsification, forged, and fraudulent evidence. However, some departures from the protocol and misconduct in clinical trials have led to tragic death of patients or healthy volunteers. These deaths included not only patients, but also normal healthy volunteers participating in the trials conducted by the world's finest medical institutions, such as Johns Hopkins University School of Medicine, University of Pennsylvania, Case Western University, and University Hospitals of Cleveland.

Example 14.4.5 The Holden-Able Case

A trial was conducted by the Case Western University and the University Hospitals of Cleveland to investigate the metabolism of the amino acids methionine and homocysteine in subjects with Alzheimer's disease and age-matched healthy controls. Methionine is sold over the counter as a nutritional supplement. On April 4, 2001, several hours of drinking a mixture of methionine and orange juice, Holden-Able, a 70-year-old matched healthy control subject in the trial became critical ill. Unfortunately, she died on May 6, 2001. The

internal investigation cannot rule out the possibility of an overdose of methionine. This example demonstrates the hidden risk in clinical trials even though the investigational product is a nutritional supplement that can be bought over the counter without prescription and the subject is a normal healthy volunteer.

Example 14.4.6 The Gene Therapy Case

The other example is the death of 18-year-old Jesse Gelsinger in a gene-transfer trial at the University of Pennsylvania for a genetically altered virus to treat an inherited liver disease (Shalala, 2000). An inspection by the U.S. FDA found the following problems in conduct and execution of the trial protocol in informed consent, patient exclusion criteria, trial stopping rule, protocol changes, and reporting adverse events. An investigation by the NIH Recombinant DNA advisory committee revealed that reporting unexpected and serious adverse events had been one of the most important protocol violations in gene therapy trials sponsored by the U.S. NIH. Before the death of Jesse Gelsinger, only 39 adverse events from gene therapy trials were reported. However, after his death, the number of reporting adverse events to the NIH Recombinant DNA advisory committee jumped to 652, a 17-fold increase. Although gene therapy is a pioneering and advanced treatment that provides real hope for some patients with previously incurable disease, principles and methodology for evaluation of efficacy and safety of gene therapy are as old as those first applied by Britain's Medical Research Council some 70 years ago.

Example 14.4.7 The Ellen Roche Case

The final example of negligence and failure of the investigator to notice and report unexpected and serious adverse events in clinical trials is the death of 24-year-old Ellen Roche. As a normal healthy volunteer and the third subject, Ellen Roche participated in a clinical trial entitled *Mechanisms of Deep Inspiration-Induced Airway Relaxation* at Johns Hopkins Asthma and Allergy Center. In this trial, hexamethonium was selected to investigate the mechanism of airway hyperresponsiveness. Hexamethonium was used to treat hypertension but was withdrawn from the U.S. market in 1972 because of ineffectiveness found by the U.S. FDA. Therefore, inhalation of hexamethonium at the time of the trial was in fact an experimental use of a non-FDA-approved drug, for which an IND should be submitted to the U.S. FDA. However, Dr. Alkis Togias, the principal investigator of the study, failed to do so. In addition, the first subject after receiving the treatment developed shortness of breath and a cough. Again, Dr. Togias did not immediately report the adverse events to the IRB. The second subject received the treatment while the first subject still had symptoms. But this subject did not develop any adverse event. However, on May 5, 2001, one day after Ms. Roche inhaled about 1 g of hexamethonium, she developed a cough. She was hospitalized on May 9 and died on June 2, 2001. On the same day that Ms. Roche was hospitalized, Dr. Togias found out that hexamethonium can have pulmonary toxic effects. A Johns Hopkins University internal investigation criticized Dr. Togias for failure to expedite reporting of the adverse event in the first subject, for failure to conduct a thorough search for previous report on pulmonary toxic effect of hexamethonium, and for failure not to delay the treatment of the next subject until the adverse events of the first subject had resolved and proper actions could be taken to protect the subsequent subjects. These examples demonstrate how important it is to follow the protocol and to report any suspected, unexpected, and serious adverse experiences and to amend the protocol to protect the trial subjects if necessary.

14.5 MONITORING, AUDIT, AND INSPECTION

The examples given in Section 14.4 clearly demonstrate that a full compliance of a clinical trial protocol is not only critical for scientific validity of the study results but also crucial for protection of human subjects who are willing to serve as experimental units in the trial. All parties involved in clinical trials, including sponsors, investigators, medical institutions, and trial subjects, should make sure that the conduct of the trial is in compliance with the approved protocol. Depending on intensity, frequency, and independence, monitoring, auditing, and inspection are performed to achieve this goal. What follows, based on the ICH E6 guideline, *Good Clinical Practice: Consolidated Guidance* (ICH, 1996), provides an introduction to the concepts, principles, and procedures regarding monitoring, auditing, and inspection of a clinical trial sponsored by a pharmaceutical company. Although they are covered from a viewpoint of the pharmaceutical industry, the following discussions can be also applied to any trials sponsored by parties other than drug companies.

According to the ICH E6 GCP guideline (ICH, 1996), monitoring of a clinical trial is defined as the act of overseeing the progress of a clinical trial, and of ensuring that is conducted, recorded, and reported in accordance with the protocol, standard operating procedures (SOPs), Good Clinical Practice, and applicable regulatory requirement(s). From this definition, the objectives of trial monitoring include the following:

- To confirm that the right and well-being of human subjects are protected
- To confirm that the conduct of the trial is in compliance with the current approved protocol/amendment(s), with GCP, and with applicable regulatory requirement(s)
- To verify that the reported trial data are accurate, complete, and verifiable from source documents

The most crucial document for a successful trial monitoring is a comprehensive and well-written standard operating procedure on trial monitoring by the sponsor and a carefully thought out monitoring plan prepared for each particular trial. Based on the objectives, design, complexity, blinding, sample size, and endpoints of the trial, the monitoring plan should describe the extent and nature of monitoring based on the guidelines given in the SOP. In general, on-site monitoring before, during, and after the trial is always necessary. Although each monitoring visit to the site is equally important, the initiation visit before the start of the trial and the close-out monitoring visit after the completion of the trial are most crucial monitoring visits. In addition, monitoring is usually more frequent in the beginning of the trial due to the trial issues arising based on the fact that investigators and their staff, patients, and even the sponsors are not familiar with the protocol.

The key personnel to the success of trial monitoring are the trial monitors from the sponsor. A trial monitor must be appropriately trained and should have scientific and/or clinical knowledge to adequately monitor the trial. Most importantly, a trial monitor should be familiar with the investigational products, the protocol, written informed consent form and any other written information to be provided to the trial subjects, the sponsor's SOPs, GCP, and applicable regulatory requirement(s). The monitor's responsibilities as defined in the ICH E6 GCP guideline include the following:

- Serving as the main line of communication between the sponsor and the investigator
- Confirming that the investigator has adequate qualifications and resources throughout the entire duration of the trial

- Verifying that the investigational product(s) are properly stored; are dispensed only to the eligible patients with adequate documentation; and their receipt, use, and return are controlled and properly documented; and that the disposition of unused investigational product(s) is in compliance with regulatory requirement(s) and the sponsor's authorized procedures
- Confirming that the investigator follows the approved protocol and approved amendment
- Verifying that written informed consent form was obtained before each subject's participation in the trial
- Ensuring that the investigator receives all documents relevant to the trial and the investigational product(s) that are necessary to conduct the trial and to comply with regulatory requirement(s)
- Ensuring that the investigator and the trial staff are adequately informed about the trial
- Confirming that the investigator is enrolling only eligible subjects
- Monitoring the subject recruitment rate
- Verifying that source data/documents and other trial records are accurate, complete, and keep up-to-date
- Confirming that the investigator provides all the required reports, notifications, applications, and submission that can identify the trial and trial subjects
- Checking the accuracy and completeness of the entries on Case Report Forms (CRF), source data/documents, and other trial-related records against each other to avoid any alteration, falsification and fabrication of the data
- Informing the investigator of any CRF entry error, omission, or illegibility
- Determining whether all adverse events are appropriately reported within the time period required by GCP, the IRB, the sponsor, and the regulatory requirement(s)
- Determining whether the investigator is maintaining the essential documents
- Communicating deviations from the protocol, SOPs, GCP, and the applicable regulatory requirements to the investigator and taking appropriate action designed to prevent recurrence of the detected deviations

The trial monitor should prepare and submit a written report to the sponsor after each on-site visit or any trial-related communication such as phone contact with the investigator. The monitoring report should include the date, site, name of the monitor, name of the investigator, and names of other personnel contacted. The content of a trial monitoring report usually contains a summary of what the monitor reviewed and the statements regarding the significant findings/facts, deviations, deficiency, conclusions, actions taken or to be taken, and/or actions recommended for compliance.

Monitoring is a key function for execution of a trial protocol. Monitors are in fact members of the sponsor's project team and are actively involved with the conduct of the trial. Therefore, the sponsor of the trial usually conducts a second-tier, independent audit by someone not involved with the trial to evaluate performance of monitors and conduct of the trial. The ICH E6 GCP guideline defines an audit as a systematic and independent examination of trial-related activities and documents to determine whether the evaluated trial-related activities were conducted, and the data were recorded, analyzed, and accurately reported according to the protocol, sponsor's SOPs, GCP, and the applicable regulatory requirement(s). In general, there are two types of audits. The first type are the routine audits that are conducted periodically during the trial. The other type are the direct audits that are usually

performed when information of serious deficiency or noncompliance of the protocol in trial conduct has been received.

Auditors appointed by the sponsor should be qualified by training and experience to conduct audits properly and should be independent of the clinical trial and any related data collection system. The sponsor should establish an independent auditing system based on the written SOPs, which specify what to audit, how to audit, the frequency of audit, and format and content of the audit reports. Then, an audit plan and procedures for a particular trial should be prepared in accordance with the SOPs and should be included in the submission along with sample size of the trial, the type and complexity of the trial, the level of risks to the trial subjects, and any identified issues. The audit reports should document any observations and findings by auditors.

Both monitoring and audits are conducted by the sponsor of the trial to ensure that the trial is conducted, recorded, and reported in full compliance with the approved protocol and its amendment(s), the sponsor's standard operating procedures, GCP, and any applicable regulatory requirement(s). Most monitoring and audits are performed before completion of the trial. On the other hand, regulatory authorities usually conduct inspections after conclusion of the trial. The purpose of the inspection is to ensure that the claims by the sponsor are valid and are derived from the data that are accurate, complete, and verifiable from the trial conducted in accordance of the approved protocol, SOPs, GCP, and regulatory requirements. The ICH E6 GCP guideline defines *inspection* as the act by a regulatory authority of conducting an official review of documents, facilities, records, and any other resources that are deemed by the authority to be related to the clinical trial and that may be located at the site of the trial, at the sponsor's and/or contract research organization's (CRO) facilities, or at other establishments deemed appropriate by the regulatory authority. Although inspections are usually conducted after completion of the trial, regulatory authorities do perform inspection during progress of the trial when serious violations of GCP or misconduct occur. For example, the U.S. FDA performed inspections in the Poisson case in Example 14.4.1, in the gene therapy case in Example 14.4.6, and in the Ellen Roche case in Example 14.4.7.

To improve the conduct and oversight of clinical research and to ensure the protection of participants in FDA-regulated clinical trials, the U.S. FDA established the Office for Good Clinical Practice (OGCP) in October 2001. The bioresearch monitoring program in the OGCP is responsible for inspection of FDA-regulated clinical trials. The purpose of the FDA's bioresearch monitoring program is to ensure the protection of research subjects and the integrity of data submitted to the FDA in support of a marketing application. The FDA's bioresearch monitoring program uses the Compliance Program Guidance Manual (CPGM) to direct the FDA's field personnel on the conduct of inspectional and investigational activities. The Compliance Program Guidance Manuals include:

- Clinical Investigators (7348.811)
- Sponsors, Contract Research Organization and Monitors (7348.810)
- Institutional Review Board (7348.809)
- *In-vivo* Bioequivalence (7348.001)
- Good Laboratory Practice (Non-Clinical Laboratories) (7348.808)

Woollen (2000b) provided a summary of results and major deficiencies found in the inspections conducted between 1992 and 1999 for clinical investigators, sponsors, contract organization and monitors, and the institutional review board. As mentioned before, the

FDA's inspections are conducted by their field personnel for the trial assigned by the reviewing center. After inspection, all establishment inspection reports (EIRs), complete with attachments, exhibits, and any related correspondence, are to be submitted in a timely fashion to the assigning center. The FDA's bioresearch monitoring program classifies the EIRs into the following three categories:

- NAI—No objectionable condition or practices were found during the inspection.
- VAI—Objectionable conditions or practices were found, but no administrative or regulatory action is recommended.
- OAI—Regulatory and/or administrative actions will be recommended.

In addition, a disqualified/restricted/assurance list for clinical investigators can be found on the FDA website: http://www.fda.gov/ora/compliance_ref/default.htm.

14.6 QUALITY ASSESSMENT OF A CLINICAL TRIAL

As emphasized many times throughout this book, the fundamental objective of a clinical trial is to make an unbiased inference with the best possible precision to scientifically answer the clinical questions with respect to a target patient population. The principles and methodologies covered and discussed in this book present the necessary tools used in a clinical trial to achieve this objective. For the same purpose, the ICH E6 GCP guideline is a set of consolidated guidelines on good clinical practice that provides an international ethical and scientific standard for designing, conducting, recording, and reporting trials that involve the participation of human subjects. After a trial is completed, a report is usually written to describe the conduct of the study, any deviation from the protocol, observations and findings obtained from the trial, and major conclusions of the trial. It follows that assessment of the quality of a clinical trial can be based on three sources: protocol, conduct of the study, and the clinical trial report.

As pointed out by Jüni et al. (2001), the quality of a clinical trial is very difficult to define and it should address the design, conduct, analysis and reporting of a clinical trial, and their clinical relevance. However, as with any research, the quality of a clinical trial depends on the validity of its results and conclusion. The validity of a clinical trial can be also classified into internal and external validity. A clinical trial is said to be *internally valid* if the differences observed between the treatment groups, apart from random error, is only due to the treatments under investigation. Internal validity is in fact an unbiased inference. All of the principles and methods described in this book are to prevent and eliminate possible bias occurring in design, conduct, analysis, and reporting of a clinical trial. *External validity* is concerned about generalizability of the result of the present study to other clinical circumstances, including different patient populations, modification of treatment regimens, or different modalities of outcomes. It follows that any quality scale for assessment of a clinical trial should have the capacity to evaluate both internal and external validity.

A lot of instruments for assessment of quality of a clinical trial have been proposed in literature. For example, Jüni et al. (2001) indicated that 39 different quality scales were identified. Most of these instruments are composite indices that combine quality information on a range of quality components for both internal and external validity into a single numeric value. Table 14.6.1 provides the quality assessment scale for evaluation of randomized controlled clinical trials proposed by Chalmers et al. (1981). This comprehensive quality

Table 14.6.1 Quality Assessment Scale for Randomized Controlled Clinical Trials Proposed by Chalmers et al. (1981)

Checklist	Possible Score
Internal Validity	
Randomization	
Randomization was blind	10
Adequacy was evaluated	3
Blinding	
Patients were blinded to treatment group	8
Physicians were blinded to treatment group	8
Patients/physicians are blinded to outcome	4
Statistician was blinded to treatment group	2
Appearance of placebo and active drug was identical	1.5
Taste of placebo and active drug was identical	1.5
Adequacy of blinding was evaluated	3
Patient attrition	
Withdrawal and reason for withdrawal were described	3
Withdrawal were handled appropriate (e.g., intention-to-treat analysis)	4
Statistical Analysis	
Statistical analysis was appropriate	4
Multiple looks at preliminary results were accounted for	3
Avoiding random error	
Number of subjects needed in trial was estimated as a prior	3
Analysis to assess baseline comparability of groups was done	3
For negative trials, statistical power of observed difference was estimated	3
External Validity	
Patients	
Description of selection of subjects was adequate	3
Description of patients screened was provided	3
Treatment	
Therapeutic regimen was defined	3
Measure of biological activity of the active therapy was made	3
Compliance with treatment was assessed	3
Other Aspects	
Additional statistical analyses	
Life table or time-series analyses was provided	2
Analysis of subgroup was appropriate	2
If indicated, regression analysis was done	2
Data Presentation	
Test statistic and <i>p</i> -value were stated	3
Confidence interval for effect was given	3
All events used as endpoints were tabulated	2
Survival curves or data sufficient to construct survival curves were provided	4
Organizational Aspects	
Starting and stopping dates of accession were provided	2
Side Effects	
Side effects were described and statistical analysis of them was done	3

Source: Chalmers et al. (1981) and Jüni et al. (2001).

Table 14.6.2 Quality Assessment Scale for Randomized Controlled Clinical Trials Proposed by Jadad et al. (1996)

Checklist	Possible Points
Randomization	
Described as randomized?	1
Allocation sequence appropriately generated?	1
Blinding	
Described as double blind?	1
Control treatment (e.g., placebo) described as indistinguishable?	1
Patient withdrawal	
Withdrawal described for each group (including the number of patients lost or excluded, along with the reasons)?	1

Source: Jadad et al. (1996) and Jüni et al. (2001).

instrument includes components on internal validity, external validity, and other trial aspects. For internal validity, it evaluates randomization, blinding, patient attrition, statistical analyses, and reduction of random error. Evaluation of external validity includes inclusion and exclusion criteria, treatments, and compliance of treatments. Other aspects consist of additional statistical analyses, data presentation, organization of the trial, and analysis of adverse events. Depending on relative importance, different points are given to each of the 30 items on the checklist and the total score is 100. A higher score indicates a better quality.

Another widely used scale for quality assessment of a clinical trial is the instrument proposed by Jadad et al. (1996). Contrary to the Chalmer's index, as shown in Table 14.6.2, the Jadad's index only addresses internal validity. The Jadad's index includes two questions for randomization, two questions for blinding, and one question for patient attrition. If the answer for each question is a yes, the score for that question is 1. It follows that the total score of the Jadad's index is 5. However, both indices as well as other composite scales suffer the same drawback. The use of the total score has a masking effect, which makes composite indices difficult to interpret. For example, two trials with equal total score based on the Chalmer's index may have totally different deficiencies in quality of the trials.

Because most clinical trials sponsored by the pharmaceutical industry are to be submitted to regulatory authorities for market applications, the quality of registration trials receives tight and close scrutiny by both the sponsor and the regulatory authorities. For example, the protocol of a registration trial must be undergone through internal review within the sponsor and external review by the IRB of medical institutions where it will be conducted. However, the trial cannot start until the regulatory authorities approve the protocol. The purpose of internal and external reviews is to make sure that appropriate procedures are specified in the protocol to protect the trial subjects and that adequate methodologies are planned in the protocol to achieve internal and external validity of the trial. During the progress of the trial, intensive monitoring and periodic audit by the sponsor are performed to ensure that the trial is conducted and recorded in accordance with the approved protocol, SOPs, GCP, and any applicable regulatory requirements. After the trial is completed, the regulatory authority may conduct inspections to examine whether the trial subjects are protected and the data generated are credible in support for a market application. Monitoring, auditing, and inspection verify that the procedures for protection of trial subjects and methodologies for internal and external validity are executed during the conduct of the trial as those specified in the protocol.

Finally, a clinical study report is written according to either the U.S. FDA guideline for the Format and Content of the Clinical and Statistical Section of New Drug Application (FDA, 1988) or the ICH E3 guideline for *Structure and Content of Clinical Study Reports* (ICH, 1996). A clinical study report documents the following:

- What had been planned in the protocol?
- What had been conducted for the trial?
- What are the deviations from the protocol and their impacts?
- What are the findings, results of analyses, and conclusions?

Some regulatory authorities such as the U.S. FDA also ask the sponsors to submit patient-level raw data of clinical trials such that the U.S. FDA reviewers can verify the results of the analyses performed by the sponsors and perform their own analyses to further confirm the validity of the claims made by the sponsor and to uncover any critical issues that are not addressed by the sponsor. In summary, assessment of the quality of the clinical trials should cover an entire spectrum, from the protocol, to the conduct of the trial, to the study report.

On the other hand, unlike registration trials, in general, the study reports of the academic clinical trials and relevant raw data are not required to submit to regulatory authorities. However, the results and important findings of academic trials are published in peer-reviewed biomedical journals. Referees and editors of most of the biomedical journals rarely can find the sufficient resources and time that the U.S. FDA reviewers have to review huge volumes of study reports and to perform their own analyses based on patient-level raw data, even though the authors of the clinical trials are willing to submit them. They, hence, have to evaluate quality of clinical trials based on the summarized information of trials from manuscripts submitted to the journals. In 1996, the Consolidated Standards for Reporting of Trials (CONSORT) Group developed the CONSORT statement as an evidence-approach to assisting improvement of the quality of reports of randomized controlled trials (Begg et al., 1996; Rennie, 1996).

Moher et al. (2001) report the results of a study to investigate whether use of the CONSORT statement is associated with improvement of the quality of reports of randomized controlled trials. They used the papers published in 1994 from the *British Medical Journal (BMJ)*, the *Lancet*, and the *Journal of the American Medical Association (JAMA)* as the pre-CONSORT baseline and performed a comparative before-and-after evaluation for the papers (post-CONSORT) published in the same journals. They also included the papers published in the same years by the *New England Journal of Medicine* as a control that did not adopt the CONSORT statement. They found that the number of CONSORT checklist items in papers of randomized controlled trials increased in all four journals in 1998. In addition, an increase of the three journals (*BMJ*, the *Lancet*, and *JAMA*) adopting the CONSORT statement was statistically significant. Furthermore, based on the Jadad's index, a statistical significant improvement in the quality score of randomized controlled trials was found in these three journals. Based on the 1996 CONSORT statement, Huwiler-Münter et al. (2002) used 25 items from the 1996 CONSORT statement as a measure of methodologic quality for clinical trials. These 25 items are given in Table 14.6.3. Papers from placebo-controlled trials published in English-language journals from 1985 to 1997 were used to examine the relationship between reporting quality and methodologic quality

Table 14.6.3 Twenty five (25) Items from the 1996 CONSORT for Evaluation of Methodological Quality

-
1. Does the title identify the study as a randomized controlled trial?
 2. Is the structural presented in a structured format?
 3. Are the objectives stated?
 4. Is the hypothesis stated?
 5. Is the study population described?
 6. Are inclusion and exclusion criteria described?
 7. Are the intervention described?
 8. Are the outcome measures described?
 9. Is a primary outcome specified?
 10. Is a minimum important different for the primary outcome reported?
 11. Are power calculation described?
 12. Is the rationale for the statistical analyses explained?
 13. Are the method for statistical analyses described?
 14. Are stopping rule described?
 15. Is the unit of randomization described?
 16. Is the method used to generate the allocation schedule described?
 17. Is the timing of assignment described?
 18. Is the number of eligible patients reported?
 19. Are prognostic variables by treatment and control group described?
 20. Is the number of receiving intervention as allocated reported for each comparison group?
 21. Are withdrawals and dropouts described for each treatment group?
 22. Are protocol deviations described for each comparison group?
 23. Is the estimated effect of the intervention on primary and secondary outcomes stated, including a point estimate and measure of precision (confidence interval)?
 24. Are the results stated in absolute numbers?
 25. Are summary data and inferential statistics presented in sufficient detail to permit alternative analyses and replication?
-

Source: Huwiler-Müntern et al. (2002).

by Huwiler-Müntern et al. (2002). They found that similar quality of reporting might hide important differences in methodologic quality. For example, the median scores on reporting quality of 33 trials using intention-to-treat analysis is 15.0, whereas it is 14.5 for 14 trials using on-treatment analysis (p -value = 0.67). Therefore, they conclude that a well-conducted trial may be reported badly and a clear distinction should be made between conduct and reporting in assessment of quality of clinical trials. Also see Devereaux et al. (2002) for a relationship between adherence of the CONSORT statement and the methodological factors in randomized controlled trials.

To overcome these shortcomings, the CONSORT group published its revised recommendations in improving the quality of reports of parallel group trials in April 2001 (Moher et al., 2001), which is given in Table 14.6.4. However, the CONSORT statement is still a measure for assessment of the quality of reporting a clinical trial and it is not for evaluating the quality of a clinical trial. A well-designed, well-conducted but poorly written report trial may not score high on the CONSORT statement. On the other hand, a biased, a poorly conducted but well reported trial may receive full credit in the quality score for reporting of trials. One should note that in addition to reporting, evaluation of the

Table 14.6.4 The CONSORT Checklist

Checklist	Descriptor
Title and Abstract	How participants were allocated to intervention?
Introduction (Background)	Scientific background and explanation of rationale
Methods	
Participants	Eligibility criteria for participants and the setting and locations where the data were collected
Interventions	Precise details of the interventions intended for each group and how and when they were actually administrated
Objectives	Specify objectives and hypotheses
Outcomes	Clearly defined primary and secondary outcomes and, when applicable, any methods used to enhance the quality of measurement
Sample Sizes	How sample size was determined and, when applicable explanation of any interim analyses and stopping rule
Randomization	
Sequence Generation	Method used to generate the random allocation sequence, including details of any restriction (e.g., blocking, stratification)
Allocation Concealment	Method used to implement the random allocation, clarifying whether the sequence was concealed until intervention was assigned
Implementation	Who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups
Blinding	Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment, if done, how the success of blinding was evaluated
Statistical Methods	Statistical methods used to compare groups for primary outcome(s); methods for additional analyses, such as subgroup analysis and adjusted analyses
Results	
Participant flow	Flow of participants through each stage (a diagram is strongly recommended). Specifically, for each group report the number of participants randomly assigned, receiving intended treatment, completing the study protocol, and analyzed for the primary outcomes. Describe protocol deviations from study as planned, together with reasons
Recruitment	Dates defining the periods of recruitment and follow-up
Baseline Data	Baseline demographics and clinical characteristics
Number Analyzed	Number of participants (denominator) in each group Included in each analysis and whether the analysis was by “intention-to-treat”. State the results in absolute numbers when feasible (e.g., 10/20, not 50%)
Outcome and estimation	For each primary and secondary outcome, a summary of results for each group, and estimated effect size and its precision (e.g., 95% confidence interval)
Ancillary analyses	Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those prespecified and those exploratory
Adverse events	All important adverse events or side effects in each intervention group
Comment	
Interpretation	Interpretation of the results, taking into account study hypotheses, sources of potential bias or imprecision, and the dangers associated with multiplicity of analyses and outcomes
Generalizability	Generalizability (external validity) of the findings
Overall Evidence	General interpretation of the results in the context of current evidence

Source: Moher et al. (2001).

quality of clinical trials include design, conduct, and analysis. Furthermore, we also believe that protection of the human subject be on the top of checklist for any instrument measuring the quality of clinical trials.

14.7 DISCUSSION

Clinical trials are scientific/biomedical experimentation for which human subjects are experimental units. The top priority for any clinical study is to protect the subjects who are willing to sacrifice themselves for advancement of science and medicine. The keys for elimination of unnecessary exposure of trial subjects to investigational intervention are (1) ethically justifiable and scientifically sound objectives and (2) a sufficient sample size to reach a valid and definitive conclusion. As indicated before, the study objectives are the most difficult task for any clinical trial. It must be based on the fundamental *uncertainty principle* (Peto and Baigent, 1998). A trial is considered only if there is substantial uncertainty about treatment effects given the assumptions that all treatments in the trial provide clinically meaningful benefits to patients. For example, an equivalence/noninferiority trial is conducted to test the hypothesis of whether the efficacy of the test treatment is equivalent or noninferior to the control given that the test treatment provides other benefits to the patients, such as a better safety profile, an easy administration route, or an improvement of quality of life. If a treatment was known to be better or do substantial harm, then the trial would not be justified. For example, it is unethical and unscientific to conduct a prospective randomized trial to verify that smoking can cause lung cancer. On the other hand, it is probably appropriate to conduct a randomized clinical trial to test the effectiveness of a school-based smoking prevention program.

Once ethically justifiable and clinically sound objectives are formulated, then the trial should have a sufficient number of patients such that a conclusion with respect to the trial objectives can be reached with a high level of confidence. Any underpowered trial not only fails to provide a concrete conclusion from the trial but also exposes trial subjects to the potential risks and harms caused by the treatments in the study. Therefore, the contributions of trial subjects are totally wasted. To plan and conduct an underpowered study is unethical too, and one should never perform a trial with inadequate power. Halpern et al. (2002) pointed out that underpowered clinical trials are ethical in only two situations:

- They are for rare diseases with a prospective plan for meta-analysis with the results with those of similar trials
- They are early phase trials in drug development, and they are adequately powered for defined purposes other than randomized treatment comparisons, such as the use of Simon's two-stage design for screening efficacy of cancer drugs

Given that the objectives are ethical and scientific and a sufficient sample size was used to provide an adequate power for the study, all principles and methodologies such as randomization, blinding, and use of appropriate control covered in this book are planned in the protocol to achieve internal validity. If the trial is conducted, recorded, analyzed, and reported in full compliance with the protocol, then an unbiased inference with respect to the trial can be made to the target patient population described in the protocol. However, 100% of accordance with the protocol rarely occurs. Therefore, the inference drawn from the trial may be

biased. The severity of the bias is a function of departures from the protocol. On the other hand, if the methods for avoiding bias are not planned in the protocol, even though the protocol is fully implemented, a biased inference is still obtained from the trial. The trial protocol determines the validity and quality that the trial can achieve. The conduct, analysis, and reporting are the processes to achieve the validity and quality defined by the protocol.

15

CLINICAL DATA MANAGEMENT

15.1 INTRODUCTION

Madison and Plaunt (2003) define clinical data management (CDM) as the process of collecting and validating clinical information with the goal of converting it into an electronic format to answer research questions and to preserve it for future scientific investigation. CDM is an integral part of the clinical trial process, which ensures the validity, quality, and integrity of data collected from trial subjects to a database system. CDM delivers a clean and high-quality database for statistical analysis and consequently enables clinical scientists to draw conclusions regarding the effectiveness, safety, and clinical benefit/risk of the drug product under investigation. An invalid and/or poor quality database may result in wrong and/or misleading conclusions regarding the drug product under investigation. Thus, the objective of the CDM process in clinical trials is not only to capture the information that the intended clinical trials are designed to capture, but also to ensure the validity, quality, and integrity of the collected data. In general, the CDM process includes (1) case report form (CRF) development; (2) database development and validation; (3) data entry, query, and correction; (4) data quality assurance; and (5) data lock, archive, and transfer. Figure 15.1.1 provides an example of biostatistics and CDM process that is commonly adopted by most pharmaceutical companies in clinical trials.

In clinical trials, it is not uncommon to encounter the following obstacles in the CDM process. First, the CDM process fails to collect useful information for addressing the scientific/clinical questions that the clinical trial intends to answer. Second, the CDM process may collect information that is irrelevant to the study objectives of the clinical trials. In addition, the quality of the collected data may be poor with missing values and/or inconsistencies across case report forms and/or study sites. As a result, the integrity of the clinical

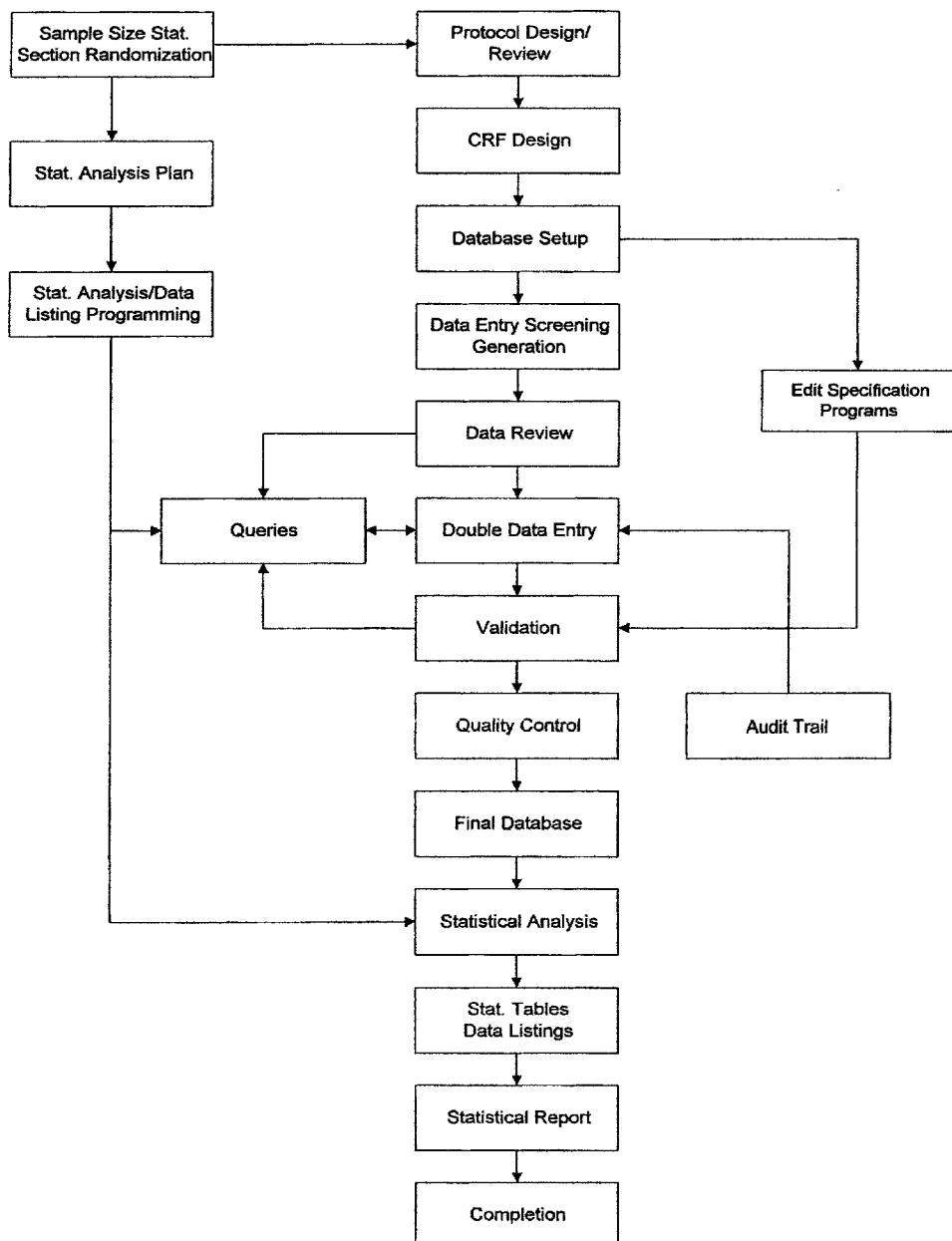


Figure 15.1.1 Sample biostatistics and clinical data management process.

trial is doubtful and the accuracy and reliability of the clinical data collected from the clinical trial is questionable. Consequently, the conclusions drawn from the analysis of the clinical data may not be accurate and reliable and hence misleading. To overcome these obstacles, the implementation of good data management practice (GDMP) is necessary. GDMP is a set of standards/procedures for assurance of the validity, quality, and integrity of clinical data collected from trial subjects to a database system. Although there are no

harmonized GDMP regulations and/or guidances/guidelines on CDM of clinical data for clinical trials, the requirements of GDMP have been inherently stated in FDA regulations (e.g., Code of Federal Regulations) and guidances and ICH guidelines. In the pharmaceutical industry, a set of standard operating procedures (SOPs), which describe the functions and individual responsibilities at various stages of the CDM process, are usually developed and implemented for compliance of GDMP.

The remaining of this chapter is organized as follows. In the next section, regulations and guidances/guidelines regarding CDM are described. Section 15.3 discusses the development of a case report form (CRF), including CRF design, review, flow, and tracking. Section 15.4 describes procedures for database development and validation. Procedures for data double entry, query, and correction are described in Section 15.5. Issues regarding data validation and data quality are presented in Section 15.6. Section 15.7 discusses practical issues in database lock, archive, and transfer. A brief discussion is given in the last section.

15.2 REGULATORY REQUIREMENTS

As indicated earlier, GDMP is the key to the success of good statistics practice (GSP), which in turn is the foundation of good clinical practice (GCP). However, there are no harmonized GDMP guidelines on data management of clinical data for clinical trials. The requirements of GDMP can be seen in regulations and/or guidances/guidelines set forth by regulatory agencies such as the Code of Federal Regulations (CFR), FDA guidances for industry, and ICH guidelines. As pointed out by Madison and Plaunt (2003), the most relevant regulations and guidances/guidelines regarding CDM process include (1) 21 CFR Part 11—*Electronic Records, Electronic Signatures* (FDA, 1997b, 2003b), (2) 21 CFR Part 312—*Investigational New Drug Application*, (3) 21 CFR Part 314—*Applications for Food and Drug Administration Approval to Market a New Drug*, (4) FDA guidance on *Bioresearch Monitoring—Compliance Program Guidance Manual* (FDA, 2001b), (5) FDA *Guidance for Industry—Computerized System Used in Clinical Trials* (FDA, 1999), and (6) ICH E6 guideline—*Consolidated Guideline for Good Clinical Practice* (ICH, 1997). In what follows, the purpose and scope of each of these relevant regulations or guidances/guidelines that directly address CDM are described.

21 CFR Part 11—Electronic Records, Electronic Signatures

21 CFR Part 11 describes the criteria under which the FDA will consider electronic records and signatures to be generally equivalent to paper records and handwritten signatures. It applies to any records required by the FDA or submitted to the FDA under agency regulations. To reinforce Part 11 compliance, FDA has published a compliance policy guide—CPG 7153.17, Enforcement Policy: 21 CFR Part 11 Electronic Records, Electronic Signatures. In addition, the FDA also published numerous draft guidance documents to assist the sponsors for Part 11 compliance. These draft guidance documents include (1) *Guidance for industry, 21 CFR Part 11; Electronic Records; Electronic Signatures Validation*, (2) *Guidance for industry, 21 CFR Part 11; Electronic Records; Electronic Signatures Glossary of Terms*, (3) *Guidance for industry, 21 CFR Part 11; Electronic Records; Electronic Signatures Time Stamps*, (4) *Guidance for industry, 21 CFR Part 11; Electronic Records; Electronic Signatures Maintenance of Electronic Records*, and (5) *Guidance for industry, 21 CFR Part 11; Electronic Records; Electronic Signatures Electronic Copies of Electronic Records*.

21 Part 11 has a significant impact on the CDM process, which has recently become the focus for GDMP in compliance with GSP/GCP. For example, 21 CFR Part 11 requires that procedures regarding creation, modification, maintenance, and transmission of records must be in place to ensure the authenticity and integrity of the records. In addition, the adopted systems must ensure that electronic records are accurately and reliably retained. 21 CFR Part 11 has specific requirements for audit trail systems to discern invalid or altered records. For electronic signatures, they must be linked to their respective electronic records to ensure that signatures cannot be transferred to falsify an electronic record. The FDA requires that systems must have the ability to generate documentation suitable for FDA inspection to verify that the requirements set forth by the 21 CFR Part 11 are met. In recent years, the compliance of 21 CFR Part 11 has become the focus of the pharmaceutical industry. Because a large portion of 21 CFR Part 11 has direct relevance to CDM, CDM has usually been identified as the top priority in the plan for 21 CFR Part 11 compliance. A typical plan for 21 CFR Part 11 compliance for CDM process usually includes gap assessment, user requirements specification, validation master plan, and tactical implementation plan. The task is implemented through a team consisting of senior experienced personnel from multiple disciplinary areas such as information technology (IT), programming, and data managers.

It should be noted that in its recent draft guidance, the FDA withdrew the draft guidance for industry 21 CFR Part 11, including *Electronic Records*, *Electronic Signatures*, and *Electronic Copies of Electronic Records*, due to the concern that some interpretations of the Part 11 requirements would (1) unnecessarily restrict the use of electronic technology in a manner that is inconsistent with the FDA's stated intent in issuing the rule, (2) significantly increase the costs of compliance to an extent that was not contemplated at the time the rule was drafted, and (3) discourage innovation and technological advances without providing a significant public health benefit (FDA, 2003a). As stated in the 2003 draft guidance, there is a narrow interpretation of scope. Under this narrow interpretation of the scope of Part 11, with respect to records required to be maintained or submitted, when persons choose to use records in electronic format in place of paper format, Part 11 would apply. Under this consideration, the FDA considers Part 11 to be applicable to the following records or signatures in electronic format. They are (1) records that are required to be maintained by predicate rules and that are maintained in electronic format in place of paper format; (2) records that are required to be maintained by predicate rules, are maintained in electronic format in addition to paper format, and are relied on to perform regulated activities; (3) records submitted to the FDA, under the predicate rules, in electronic format; and (4) electronic signatures that are intended to be the equivalent to handwritten signatures, initials, and other general signings required by predicate rules. The FDA intends to exercise enforcement discretion regarding the specific Part 11 requirements for validation, audit trail, legacy systems, copies of records, and record retention.

21 CFR Parts 312 and 314

As indicated earlier, 21 CFR Parts 312 and 314 contain regulations for submitting an investigational new drug application (IND) and a new drug application (NDA) or an abbreviated new drug application (ANDA). Both regulations have direct relevance to CDM, especially the handling of case record form (CRF). Specifically, Part 312.63 (b) covers requirements for investigator's record-keeping and record retention. The investigators are required to keep adequate case histories that record all data pertinent to the investigation on each individual administered an investigational drug or used as a control in the investigation. On the other

hand, Parts 314.50 (f) (1) and (f) (2) pertain to the inclusion of CRFs and tabulations in an application. The FDA requires that full case report tabulations from each adequate and well-controlled study (phases 2 and 3 studies) and from clinical pharmacology studies (phase 1 studies). Safety data tabulations are required from other clinical studies. These regulations have a direct impact on the CDM process.

FDA Guidance on Bioresearch Monitoring—Compliance Program Guidance Manual

The Bioresearch Monitoring guidance provides specific instructions to sponsors, clinical investigators, laboratories, and institutional review boards regarding FDA inspection for compliance of requirements set forth to ensure the quality and integrity of clinical data submitted to the FDA. This guidance provides specific instructions for FDA inspections on data collection and handling, including the review of data tabulations for each subject, the review of the sponsor's and/or contract research organization's (CRO) data collection, and the handling of SOPs. This guidance also covers requirements for automated entry of clinical data in compliance with 21 CFR Part 11 described earlier. As a result, this guidance also has a direct impact on the CDM process. In particular, program 7348.810 of the guidance is directly applicable to the CDM process.

FDA Guidance for Industry—Computerized System Used in Clinical Trials

This guidance provides general principles that are to be followed when computerized systems are used to create, modify, maintain, archive, retrieve, or transmit clinical data intended for submission to the agency. Although the primary focus of this guidance is computerized systems used at clinical sites for data collection, it can be applied to sponsors and CROs. It should be noted that this guidance addresses requirements similar to those described in 21 CFR Part 11 for electronic records and electronic signatures.

ICH E6 Guideline—Consolidated Guideline for Good Clinical Practice

ICH E6 *Guidelines for Good Clinical Practice* sets forth a tripartite standard for the conduct of clinical trials among the United States, European Union, and Japan. It covers preparation, monitoring, reporting, and archiving of clinical trials. The ICH E6 guideline indicates that quality control should be applied to each stage of data handling to ensure that all data are reliable and have been processed correctly. In addition, any changes or corrections to a CRF should be dated, initialed, and explained (if necessary) and should not obscure the original entry. This applies to both written and electronic changes or corrections. In any way, an audit trail should be maintained. The ICH E6 guideline suggests that certain quality control procedures should be employed during data management processing and the procedures should be available for audit.

Most recently, a committee of the Society for Clinical Data Management (SCDM) proposed a draft good clinical data management practice (SCDM, 2000). The draft proposal provides comprehensive details regarding GDMP. Fong (2001) indicated that GDMP refers to the compliance with ICH, GCP and FDA guidelines on CDM for clinical trials. Fong (2001) identified the three most important concepts for GDMP as (1) the implementation of quality control procedures, (2) an audit trail, and (3) quality quantification of the final database.

15.3 DEVELOPMENT OF CASE REPORT FORMS

The CDM process begins with the development of case report form (CRF), which takes place at an early stage of protocol development. The CRF should be designed to capture correct information in an effective way. The CRF process includes procedures for handling CRFs and CRF flow and tracking, which is an important factor for the success of the CDM process. In what follows, issues in CRF development, procedures for handling CRF flow, and the CRF tracking system will be briefly outlined.

CRF Design

The CRF is the most commonly used data collection instrument or tool in clinical trials. The design of CRFs for a given clinical trial requires collaboration of personnel from the clinical, data management, statistical, and programming areas. Madison and Plaunt (2003) recommended that the following principles be considered when designing CRFs for an intended clinical trial: (1) The CRF should be designed to capture all data required per the study protocol, (2) the CRF should be designed to collect data elements in standardized format, (3) data elements should be captured on the CRF in a fashion that ensures that data are suitable for summarization and analysis, (4) data elements planned to be transcribed to the CRF from source documents should be organized and formatted on the CRF to reduce the possibility of transcription error and to facilitate subsequent comparison to source documents, (5) ease of completion for the investigator and study coordinator is key to accurate and timely CRF completion, and (6) redundant data elements within the CRF should be avoided and unnecessary data should not be collected. It is essential that all and only the necessary data be collected. CRF should be designed in a way that promotes simple database design, data capturing, and data validation. A typical approach for development of CRF is the so-called *backwards* approach. The backwards approach to CRF design starts with assessing the type and format in which information will be presented in the study report. Information is then tracked backwards, and the CRF is designed to capture data in a manner that allows them to be easily converted or directly placed in the designed tables or figures (Spilker, 1991). This approach requires the input of the trial/project statistician/programmers and ensures that the CRF design meets the requirements of the users of the database.

As pointed out by Grobler et al. (2001), the trial/project statistician/programmer should be involved in the development of CRF design by providing the following input: (1) Find the best balance between effective data collection and structuring the CRF to facilitate data entry, (2) design a CRF in such a way that the resulting datasets would be programmable with the least amount of data manipulation, (3) collect the minimum amount of necessary data, (4) data should be collected in a manner that facilitates data analysis, (5) calculated data should not be recorded in the CRF, but all data needed to perform the calculation should be included in the CRF, and (6) multiple clinical trials conducted for one submission should have consistent databases. During the development of CRF, it is essential to discuss the developed CRF with the relevant personnel who will be involved in the process after data have been collected. In practice, it would be beneficial if the statistician is consulted and given the chance to review a CRF regardless of whether the statistician is responsible for the design of a CRF. Data manager and statistician/programmer should have an opportunity to evaluate the CRF to ensure that the data collected are in proper format for entry and data analysis.

Table 15.3.1 General Principles for Filling Out CRF

-
1. Print neatly and legibly.
 2. Make all entries on the CRF.
 3. Write in black ink and press firmly.
 4. Do not write in shaded areas.
 5. Avoid use of abbreviations and acronyms; use only abbreviations in standard medical use.
 6. Use only subject number and initials, not the full name and chart number of the subject.
 7. Use the following abbreviations for missing data: NA (not available/not applicable), ND (not done), UNK (unknown).
 8. Comments must be clear, concise, and in language specified.
 9. Do not write data on page margins or outside of allocated spaces.
 10. Record dates as specified.
 11. Record time in 24-hour clock format.
 12. Check all visit dates in chronological order.
-

A set of well-developed CRFs is not completed without a guide as to how to fill out the CRFs. In clinical trials, errors/mistakes are inevitable when filling the CRFs regardless of the design of the CRFs. Ju (2002) proposed some general principles for filling out well-designed CRFs in clinical trials. These general principles as listed in Table 15.3.1 are helpful in reducing the number of queries and maintaining the quality and integrity of the data captured. In addition, it is also suggested that the following principles be applied when making corrections: (1) draw line(s) through the incorrect entry, (2) do not write over or erase an incorrect entry or recopy the original page, (3) do not use correction materials on the CRFs, (4) write the correct data nearby, and (5) date and initial the corrections. In practice, it is suggested that these commonly encountered errors and/or mistakes be taken into consideration in CRF design during the development of CRF.

CRF Flow and Tracking

In clinical trials, CRF flow and tracking is important in the data management process to maintain the integrity of the clinical trial. Prior to the commencement of a clinical trial, CRFs with a guide to filling out CRFs should be shipped to each study site. At the initiation investigator meeting, the CRFs will be reviewed. Any issues regarding the capture of the data using the developed CRFs will be discussed and the resulting changes will be made prior to the conduct of the clinical trial. During the conduct of the clinical trial, completed CRFs are usually reviewed for accuracy, completeness, and consistency with the study protocol by the clinical monitor and/or clinical research associate before they are forwarded to the data management group for data entry sequentially.

Upon receipt of completed CRFs, the page numbers or other unique identifiers of the CRFs, including laboratory or diagnostic test forms, should be recorded on a CRF tracking sheet such as a case report form transmission form. Table 15.3.2 provides a sample case report form transmission form, which is commonly used in the pharmaceutical companies. In addition, a case report form log sheet should be used to document medical review and data entry. The original CRF is usually separated from the working copy of the CRF. The original CRF should be filed in the Central Clinical Trial File, whereas the working copy and the accompanying log sheet should be forwarded to the data

Table 15.3.2 Sample Case Report Form (CRF) Transmission Form

Protocol Number:	Drug Name:	Sponsor Protocol Number:									
Center Number:	Investigator's Name:										
Please list the subject numbers and pages (submitter). Acknowledge receipt of CRFs by ticking the appropriate boxes (recipient).											
Subject No./Initial	Received	Subject No./Initial	Received	Subject No./Initial	Received	Subject No./Initial	Received	Subject No./Initial	Received	Subject No./Initial	Received
1	<input type="checkbox"/>	11	<input type="checkbox"/>	21	<input type="checkbox"/>	31	<input type="checkbox"/>	41	<input type="checkbox"/>	51	<input type="checkbox"/>
2	<input type="checkbox"/>	12	<input type="checkbox"/>	22	<input type="checkbox"/>	32	<input type="checkbox"/>	42	<input type="checkbox"/>	52	<input type="checkbox"/>
3	<input type="checkbox"/>	13	<input type="checkbox"/>	23	<input type="checkbox"/>	33	<input type="checkbox"/>	43	<input type="checkbox"/>	53	<input type="checkbox"/>
4	<input type="checkbox"/>	14	<input type="checkbox"/>	24	<input type="checkbox"/>	34	<input type="checkbox"/>	44	<input type="checkbox"/>	54	<input type="checkbox"/>
5	<input type="checkbox"/>	15	<input type="checkbox"/>	25	<input type="checkbox"/>	35	<input type="checkbox"/>	45	<input type="checkbox"/>	55	<input type="checkbox"/>
6	<input type="checkbox"/>	16	<input type="checkbox"/>	26	<input type="checkbox"/>	36	<input type="checkbox"/>	46	<input type="checkbox"/>	56	<input type="checkbox"/>
7	<input type="checkbox"/>	17	<input type="checkbox"/>	27	<input type="checkbox"/>	37	<input type="checkbox"/>	47	<input type="checkbox"/>	57	<input type="checkbox"/>
8	<input type="checkbox"/>	18	<input type="checkbox"/>	28	<input type="checkbox"/>	38	<input type="checkbox"/>	48	<input type="checkbox"/>	58	<input type="checkbox"/>
9	<input type="checkbox"/>	19	<input type="checkbox"/>	29	<input type="checkbox"/>	39	<input type="checkbox"/>	49	<input type="checkbox"/>	59	<input type="checkbox"/>
10	<input type="checkbox"/>	20	<input type="checkbox"/>	30	<input type="checkbox"/>	40	<input type="checkbox"/>	50	<input type="checkbox"/>	60	<input type="checkbox"/>

Total number of CRFs: _____

Submitter: _____	Name _____	Date _____	Signature _____
Recipient: _____	Name _____	Date _____	Signature _____

management group. In addition, all CRFs should be cross-referenced against the accompanying log sheet.

During the conduct of clinical trials, errors and mistakes inevitably occur regardless which CRFs are used in clinical trials. Commonly seen mistakes in CRFs are as follows: (1) There are no page numbers, (2) dates are inconsistent with charts, (3) visit dates are not in chronological order, (4) there are no investigator's signature, (5) laboratory units are either incorrect or inconsistent, (6) inconsistent between fields (e.g., action taken with medication recorded in adverse events but not shown in concomitant medication), (7) missing data (e.g., should fill in N/A), (8) error erased and/or correction made without initial or signature, (9) inconsistencies across CRF pages, and (10) calculation error.

In the past few decades, the standardization of CRFs has become a popular topic. The purpose of standardized CRFs is multifold. First, it is to ensure the consistency of data capture. Second, it is to reduce the time and effort for development of CRFs when new studies are planned. However, it is a difficult task to standardize CRFs across therapeutic areas due to differences in the nature of disease, study design, and dose regimen.

For multinational clinical trials, translation of the developed CRFs has become an issue for capturing consistent information across different countries with different languages. The perception of the CRFs may be different due to differences in culture and medical practice in different countries. A typical approach is to adopt the method of translation and backtranslation to validate the translated CRFs. For this purpose, a small scale trial may be required to be conducted with the translated CRFs for validation of the translated CRFs.

15.4 DATABASE DEVELOPMENT

As indicated by Grobler et al. (2001), a database should be designed to facilitate data entry and the extraction of data for analysis. Database development includes database design (or setup) and database edit check specifications, which are briefly outlined below.

Database Design

In practice, for a given clinical trial, to facilitate data entry and the extraction of data for analysis, a protocol-specific database is set up using standard templates (e.g., modules and format libraries or data dictionaries) where available. The use of standard templates enhances the efficiency of the database development process and facilitates subsequent aggregation of the data. The steps in developing a protocol-specific database are illustrated in Figure 15.4.1 (reproduced from Fig. 1 of Madison and Plaunt, 2003). As it can be seen from Figure 15.4.1, once the applicable standard templates are identified, the protocol-specific database can be built by creating the following associated structures of (1) data entry screens, which are identical to CRFs; (2) test data; (3) derived variables; (4) data validation routines; and (5) audit trail. Prior to the generation of screens for data entry, CRFs will be reviewed for code requirements. Coding instructions will then be generated and reviewed by the data manager and the project statistician to ensure that all information needed for analysis has been captured from the CRFs. Upon the completion of data entry or at certain percentage of completion of data entry, data may be extracted for a test run of analysis. Additional variables to be derived from

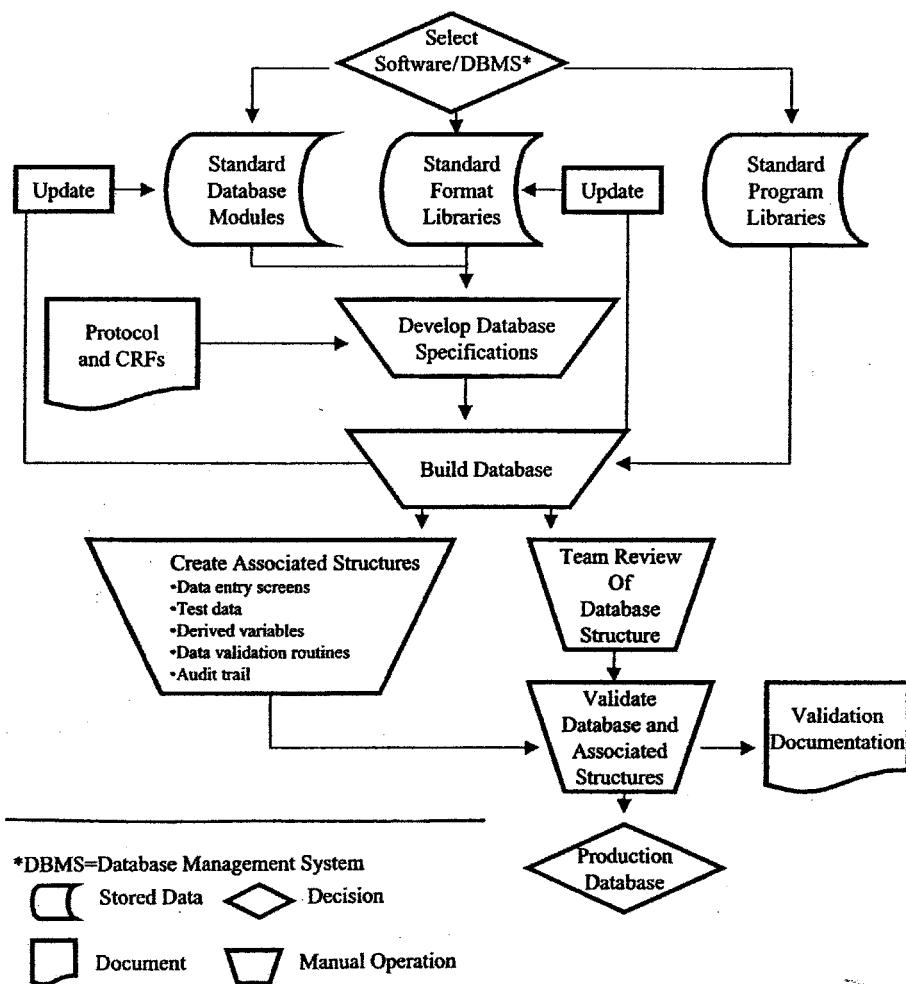


Figure 15.4.1 Sample database development process. (Source: Madison and Plaunt, 2003)

collected data may be identified by the project statistician. The derived variables (or derived data sets), derived from the database according to the statistical analysis plan, are usually ready for analysis. For a given clinical study, in some cases, it is not uncommon to have more than one database due to planned interim analyses and/or pharmacokinetic/pharmacodynamic analyses. In practice, a data validation plan is usually developed during the database development process to ensure the validity, quality, and integrity of the data captured from subjects in the clinical trial. Audit trails are required to document any changes that have occurred during the database development process.

Data Edit Check Specifications

To ensure completeness and consistency of a database, data validation or plausibility checks are usually programmed and run on a database. In small databases, or parts of databases that cannot be validated through a program such as consistency of text data, validation checks

may also be done manually. Incomplete or inconsistent data are queried with the source data at a study sites or laboratories. Once query resolutions have been received, the database is updated accordingly. Any such updates must be documented in an audit trail. In practice, it is suggested that the trial/project biostatistician should be involved in determining the consistency check specifications. The following should be considered when reviewing the consistency check specifications: (1) missing or incomplete data, (2) range checks included to detect potentially invalid data, (3) within-subject consistency checks, and (4) manual checks whenever necessary.

Edit check specifications are usually developed according to CRFs and the study protocol. Variables and fields with a high risk of potential errors on the CRFs should be included. Edit checks across CRFs should also be included in the edit check specification to capture any inconsistencies across CRFs during data entry. Queries should be generated whenever inconsistencies are detected.

Edit check specifications are important in database development not only to identify invalid data but also to capture inconsistencies in data collected from the clinical trial. Consequently, it is an integral part of database development to ensure the validity and quality of the data collected from the clinical study. For example, if the adverse event dataset denotes that a subject was withdrawn due to an adverse event, does the reason for subject withdrawal dataset indicate withdrawal due to an adverse event? For another example, if there is an adverse event indicating that medication was prescribed to treat the condition, is there a corresponding entry in the concomitant medication dataset that is appropriate according to the indication and period of administration?

It is the responsibility of the statistician to ensure that the design of the database adheres to the requirements of the programming team and that there is a consistency between the database across trials involving the same treatment, to ease the pooling the data across studies. This involves reviewing at least the following: (1) the proposed naming, formatting, and labeling of variables and datasets; (2) details of the coding systems to be used; (3) the contents of datasets, to minimize data manipulation in extracting data for analysis and to ensure that the data are captured appropriately; and (4) the proposed design and format of data to be downloaded or transferred for statistical analysis. A statistician should review these characteristics to ensure that they are consistent with the anticipated presentation of the data in tables and figures. It must be remembered that a poorly designed CRF cannot be overcome by good database design.

15.5 DATA ENTRY, QUERY, AND CORRECTION

Procedures for data entry/verification, data query, and data correction/validation play an important role in the CDM process to ensure the validity, quality, and integrity of clinical data collected from trial subjects. These procedures are briefly described below.

Data Entry and Verification

Upon the receipt of CRFs, a working copy of the original CRFs will be made and reviewed by the data manager prior to entry. Data from the working copy of the CRFs will be entered into the database. This is normally accomplished by double data entry (i.e., entry by two different entry clerks). CRFs will be initialed, dated, and stamped *entered* after the initial entry. A second entry will occur after the initial entry. The two files are then checked for differences by

running a verification program that compares the two files. If the records do not match, discrepancies are printed out and corrections are made. The process is repeated until such a time that each file is an exact replica of the other. Any missing or inconsistent information will be entered on the Query Form. A sample Query Form that is commonly used in clinical trials is given in Table 15.5.1. After the information has been verified, i.e., there are no inconsistencies between the initial and the second entries, the CRFs will be initialed, dated, and stamped *verified*. Verification ensures that data entered are actually on the CRFs. Validation checks will then be done using built-in edit check programs in the database developed by the data manager or by running SAS programs written by the project statistician.

Errors inevitably occur during data entry. Yamuah (2001) indicated that the most common errors that usually occur during data entry include (1) typographical errors, (2) copying errors, (3) coding errors, and (4) range errors. For typographical errors, this usually happens when someone is typing very fast. Copying errors usually occur as a result of poorly filled-in CRFs with handwriting that is not very legible. A typical example is that data entry clerks cannot differentiate between a 0 (zero) and an O (letter). Coding errors can originate either from the personnel filling the CRFs with given codes or data entry clerks making a mistake with given codes to be assigned to items on the CRFs. A typical example is the assignment of ethnicity code. Range errors occur where lower and/or upper limits of known values are exceeded when typing. For example, a body height of 8 feet is obviously impossible for a regular subject.

Note that the data quality of the resulting database is inversely proportional to the chance of having the same data entry error in a field by two persons, which is hopefully negligible. McFadden (1998) indicated that the overall error rate from double data entry can be as low as 0.001%. Double data entry in clinical trials is, however, only considered when CRFs are used and its cost-effectiveness has been debated (see, e.g., Gibson et al., 1994; Day et al., 1998). Although logic checks should be routinely done in all trials, visual data verification and double data entry are not mandatory. Certain data quality inspections in addition to logic checks on the database are, however, highly desirable to quantify the data quality level whenever CRFs or eCRFs are used. Statistical methods utilizing sampling techniques to estimate the error rate are available. However, whatever procedures are adopted, the estimation method of the error rate should be properly chosen and clearly documented in order to enable accurate data quality quantification (King and Lashley, 2000).

Data Query

Any errors, omissions, or items requiring clarification or changes to CRF detected during the data entry and verification process, by computer edits or during the data analysis, will be noted on the Query Form and forwarded to the clinical monitor for processing. Data queries may be generated at any time during the CRF review, data entry and verification, and data analyses by the data management personnel and/or trial/project statistician. Data clarifications not requiring CRF alterations will also be entered on the Query Form, where they will be resolved by the designated monitor. Data queries, which require an alteration to the CRF, will be noted on the Query Form. Obvious resolutions to queries (e.g., date errors and transposition of laboratory values) may be made by the data manager on the Query Form and sent to the designated monitor (or investigator) for confirmation. Query resolutions that must be obtained by the monitor should be conducted either via telephone contact with the investigator or through site visit.

Table 15.5.1 Sample Data Query Form

Protocol Number:	Drug Name:	Page ____ of ____	
Patient No./Initials:	/	Investigator/Center: _____ / _____	
Query No.	CRF Page No. or Identified	Question	Resolution
			Clinical Monitor Investigator
FOR INTERNAL USE ONLY: Original CRFs corrected and database revised to reflect corrections.			
		REQUESTED BY: _____ / _____	____ / ____ Date
		PROCESSED BY: _____ / _____	____ / ____ Clinical Monitor Date

Investigators will be asked to sign and date the form and retain one copy of the completed Query Form with their copy of the CRF. The original Query Form will be returned to the individual initiating the query for review. All confirmations and resolutions should be forwarded to the data management group.

Data Corrections/Validation

All subsequent changes to the database should be made based on information provided on the Query Forms. Upon completion of data entry, original Query Forms should be dated, initialed, and stamped *Corrected* to indicate that the entry of the data update is completed. Database finalization will begin after all queries have been addressed and necessary modifications have been made to the database. Note that working copies of the CRFs are necessarily updated during the process of data entry and verification, query, and corrections.

Validation ensures that data entered into the database are valid according to some criteria usually arrived at by an expert in that area. It is important to note that verified data are not necessarily accurate data. If data are invalid from the field that will be verified, correctly using double entry but will still remain invalid. If at any stages errors are detected, CRFs should be returned to the field for clarification or revisited depending on the nature of the problem.

A database must be reviewed to ensure that it is consistent with the information recorded in the CRF and that any data that were received electronically have been correctly imported into the database. In conducting the quality control of a database, it is necessary to determine which of the efficacy and safety database items are deemed to be critical and noncritical items, respectively. A statistician should be responsible for determining the critical items and the population to be audited in accordance with the objectives of the clinical trial. Procedures for checking the various types of data should be specified and reviewed for appropriateness. These procedures may include, for example, a 100% check of all data items considered to be critical against the CRF. An acceptable database error rate should be determined, and acceptable and nonacceptable errors should be defined in conjunction with the project statistician.

15.6 DATA VALIDATION AND QUALITY

Data validation is the cleaning of trial data after they are entered into a computer database in order to ensure that they attained a reasonable quality level. Errors occur whenever a true value is not correctly represented in a certain format. This could be a discrepancy between source data and the CRFs, source data and an electronic data captured database, or the CRFs and the database, depending on the data acquisition model (Fong, 2001). Detection of the former two sources of error relies on monitoring visits when source data certification is performed (Lau, 2000). The third source of errors can be identified by conducting data quality inspections on the database.

Although data quality assurance is highly prioritized in clinical trials, perfect data are difficult if not impossible to obtain. It has been recognized that high-quality data are used to arrive at the same conclusion as perfect data (Rasmussen, 2000). An acceptable data quality level is one that can be considered for providing substantial evidence regarding the efficacy and safety of the study medicine. However, an acceptable data quality level has not

been defined and there are no regulatory guidelines on an acceptable level of data quality. This was perhaps due to the diversity in industrial practices. Nevertheless, it is critical to quantify data quality by identifying errors from different sources, prior to any conclusions being drawn from the statistical analysis.

Data quality may be measured by the error rate defined as the number of errors divided by the total number of data. Quality can be attributed to all variables, or to a group of variables whose quality is deemed to be critical to the final conclusions. During database inspection, the error rate can be simply estimated by the number of errors found divided by the number of data inspected. The choice of an acceptable error rate for a database varies in the industry, but a popular choice was 0.5% overall, 0% to 0.1% for critical variables, and 0.2% to 1.0% for noncritical variables (see, e.g., SCDM, 2000; Shea, 2000). The methods of estimating the error rate should be documented in the Data Management Master File. Note that an estimated error rate obtained as 5 errors out of 1000 fields inspected bears a different precision to that obtained as 50 errors out of 10,000 fields inspected. The latter is a more precise estimate. There are different data quality inspection procedures for tracking database errors. The common types are logic checks, visual data verification, double data entry, or a combination of these methods. Logic checks, as described in the previous section of edit check specifications, are trial-specific procedures to determine the status of data that are not logically sound.

15.7 DATABASE LOCK, ARCHIVE, AND TRANSFER

Database Lock and Archive

When the database is complete, i.e., there are no outstanding queries, signatures of the responsible individuals are required to *finalize* (or *lock*) the database. A sample Database Lock Form, which requires signatures from the project statistician, the data manager, and the clinical monitor, is given in Table 15.7.1.

After the database lock, a written approval from senior management (e.g., Head of Clinical Operation and Head of Biostatistics and Data Management) is required to initiate any changes to a finalized database. The final locked database should remain active in the system for at least three months before it is archived. An archived database can be retrieved within two working days.

Database Transfer

Database transfer at a specific format can be done per request, e.g., from CRO to a sponsor. Data transfers are usually accomplished based on three steps of (1) Data Transfer Request Form, (2) test data transfer, and (3) data transfer procedure.

First, the requestor should fill out a Data Transfer Request Form. Table 15.7.2 provides a sample Data Transfer Request Form. Based on the information on the Data Transfer Form, programmer, data manager, and statistician will develop the data transfer specifications. The specifications are necessary to (1) ensure that the scope, schedule, and frequency of the data transfers are clearly defined; (2) make sure that the client's hardware/software are compatible with data transfers; and (3) obtain and finalize clearly defined data transfer specifications, including data structures and system requirements. Prior to the actual data transfer, a test run for data transfer is usually recommended. The purpose of performing a test run for data

Table 15.7.1 Sample Final Database Lock Sheet

Protocol #:
Protocol title:
Sponsor:

Final Database Reviewed:

Completed by: _____ Date: _____

Reviewed by: _____ Date: _____

Final Database Locked:

Project Manager: _____ Date: _____

Data Manager: _____ Date: _____

Clinical Monitor: _____ Date: _____

transfer is to identify and correct any structural errors and inconsistencies that may occur prior to the first scheduled data transfer.

Data transfer can then be completed by the following procedure:

1. Develop programs when the data specifications are available and at least five patients have been entered into the database.
2. Communicate with the Project Team for any data handling issues.
3. Perform a quality control check on the database.
4. Make sure that all data involved in the data transfer are clean if clean data are expected.
 - a. Make sure that all data involved in the data transfer are appropriately coded if coding is to be done. Forward any uncoded data to the Project Team for clarification and/or review.
 - b. Make sure that the database contains the required number of patients.
 - c. Make sure that all of the data streams are present as outlined in the data transfer specifications.
 - d. Make sure that only properly validated data are transferred to the client.
 - e. Notify the programmer(s) when the database is ready for a data transfer run after database validation.
5. Generate permanent SAS data sets for file transfers.

Table 15.7.2 Sample Data Transfer Request Form

To:	From:
Copy:	Date:
Re:	

Account: _____

Library: _____

Directory: _____

Files: _____

- Data will be ready on _____ / _____ / _____.
- Data is promised to sponsor on _____ / _____ / _____.
- Department to create cover memo: _____
Cover memo will be copied to: 1. _____ 2. _____ 3. _____
4. _____ 5. _____ 6. _____
- Database Administration: _____

Address: _____

- Schedule of Transfer:

- Method of transport

<input type="checkbox"/> Overnight carrier	<input type="checkbox"/> Courier	<input type="checkbox"/> Other
--	----------------------------------	--------------------------------

6. Perform program checks and data review during program development.
7. If the client requests flat file data transfer, create ASCII files from the corresponding permanent SAS data sets.
8. Make sure all forms and checklists related to the data transfer are signed off.
9. Make sure all documents and printouts related to data transfer are filed in a central master file.

Note that it is suggested that data transfers should be performed based on programming standards as specified in *Good Programming Practice* (GPP) (see, e.g., Yam, 2003) and follow requirements that are specified in 21 CFR Part 11 for electronic records and electronic signatures.

15.8 DISCUSSION

Data Management Files

For a given clinical study, it is important to maintain a *Master Data Management File* that contains critical information regarding the clinical study. A complete Master Data Management File should include (1) a copy of a final approved study protocol with amendments, if any; (2) a copy of blank CRFs and a Completion Guide to CRFs; (3) coding instructions and edit check specifications; (4) laboratory normals from each study sites; (5) a data validation plan; (6) resolved queries forms and outstanding queries, and (7) correspondences regarding the study. The purpose of the master data management file is to maintain key documents at each critical stage of the clinical data management process for GDMP in compliance with GCP to ensure not only the quality of the data collected from the study but also the integrity of the clinical trial.

Electronic Data Capture

In recent years, electronic data capture (EDC) has become an important topic in clinical trials data management since it was introduced in an NIH-sponsored kidney transplant histocompatibility study in early 1970. A number of EDC systems have been developed since 1985. However, many of them are not successful due to the issue of data quality. Helms et al. (2001) proposed to measure data quality by estimating error rates for specific sets of data. An error rate is defined for a database subset (a specified set of data values) not the entire database. The specified set of data values is often limited to certain types of variables (e.g., excluding long text fields and codes). The specified set of data values must be definitive (by applying the specifications) so that a computer program is able to determine precisely whether a specific data value is in the specified set of data values. The true error rate for a specified set of data values is then defined as the ratio between the number of incorrect data values in the specified set of data values and the total number of data values in the specified set of data values. A random sample can then be drawn using the method of simple random sampling to estimate the true error rate. For example, a simple random sampling may be applied to randomly select a subset of patients. Each patient is considered a cluster. Within each selected patient's data, one may use simple random sampling to select a random subset of visits (subclusters). Note that in this case, the random sample consists of all data fields of specified types within each selected subcluster.

As the Internet has become more popular, some EDC systems were designed to utilize the Internet as a means of transmitting data from study sites to a central location. In 1999, the FDA published a guidance on *Computerized Systems Used in Clinical Trials* to assist sponsors in compliance with regulatory requirements of clinical trials data management when using EDC. In practice, an EDC Internet-based system has the advantage of reducing the time and effort required for clinical trials data management. When the same procedures and standards are used to estimate EDC Internet-based system error rates that are used for estimating error rates of traditional paper-based CRF systems, Helms et al. (2001) indicated that the EDC Internet-based system has essentially zero error rates when capturing data electronically. Note that his conclusion applies only to electronically captured data in which the initial data record is on a computer disk.

Interactive Voice Randomization System (IVRS)

For management of clinical trials, the use of an interactive voice randomization system (IVRS) for patient randomization and drug management has become very popular, especially for multinational clinical trials. The IVRS is an application based on advanced computer telephone technology. It has the capability of functioning without human intervention, which makes worldwide access possible. As a result, it is an ideal tool for central randomization and drug management in clinical research. IVRS has the ability to centrally randomize the patients so that treatment balance can be achieved across all study sites. In addition, the IVRS technology can also provide clinical researchers with a new and cost-effective method of collecting real-time quality data and for providing critical monitoring information at any time and any place where a touch-tone phone is available.

As indicated in Chapter 4, the randomization schedule is used for drug management, including packaging, shipping, and dispensing in clinical trials. The traditional approach is to provide a randomization schedule to the Department of Drug Supply for generation of labels for packaging. Blocks of drugs are then shipped to each study site according to respective randomization schedules. In practice, however, some sites may have relatively slow enrollment as compared to others. As a result, the management of drug supply for reduction of drug wastage is a challenge to clinical researchers. IVRS, on the other hand, is more flexible. One of the advantages is its ability to separate patient numbers from the randomization codes and to match them dynamically at time of randomization for treatment balance across all study sites. The randomization codes serve as a bridge between the treatment assignment and any patient numbering system. IVRS is applicable to most commonly used randomization methods, as described in Chapter 4. In addition to the advantage of central randomization, one of the major attractive advantages is that IVRS has the capability of collecting real-time quality data. For example, if the protocol calls for the collection of patient diary data, it is always a headache to clinical data due to inevitable errors in diary cards and/or missing values. Using IVRS, quality data can be obtained directly from the patient. In addition, IVRS can also call and remind the patient to enter the diary data to avoid missing data. As a result, a high patient compliance for the use of IVRS is expected. IVRS is also a useful monitoring tool for providing information such as patient enrollment and withdrawal status. This information would be very helpful for strategy planning. In some studies for long-term safety, IVRS can collect real-time crucial safety information such as serious adverse events and mortality data.

Because IVRS can collect real-time quality data during the conduct of the trial, it is very useful when there are planned interim analyses. In practice, however, for good clinical practice, it is suggested that a standard operating procedure (SOP) for the process of the reconciliation between the IVRS database and the database developed based on CRFs should be well documented. More details regarding the use of IVRS in the CDM process can be found in Chen (2003).

Global Database and System

In recent years, multinational, multicenter clinical trials have become very popular for clinical evaluation of the effectiveness and safety of experimental drugs under investigation. As a result, the standardization of databases across different countries with potential differences in culture and medical practices is necessary in order to provide a consistent database for final statistical analysis. Hsuan and Genyn (2003) indicated that a global database and system is necessarily implemented for achieving standardization of databases across different countries. A successful global system requires both good system development methodology and good process. The main objective of a global clinical system is to shorten the drug development time while improving the quality of clinical data, analysis, summary, and reports. The global clinical database and system will achieve the benefits of improved submission quality, synergy among countries (sites), shared resources among countries (sites), shortened submission preparation time, shortened response time for regulatory agency inquiries, and electronic data review internally and by the regulatory agencies (Hsuan and Genyn, 2003).

Role of Statistician in Data Management Process

The statistical results from a clinical trial rely on the accuracy and completeness of the database used to generate the results. The quality of the database depends on the quality of the data collection methods and the data management procedures. As part of the data management-biostatistics team, a statistician can ensure that a protocol is written in such a manner as to enhance data management. A statistician's input into the subsequent design of the case report form can facilitate the collection of appropriate and complete data. A statistician can contribute to the accuracy and validity of a database by providing into the design both the database structure and data plausibility checks and the quality control procedures performed to evaluate the integrity of a database. Finally, performing a statistical review on the database prior to database lock can detect anomalies that would otherwise only show up during analysis (Grobler et al., 2001).

As pointed out by Grobler et al. (2001), the following statistical reviews of a database are helpful to ensure the validity, quality, and integrity of the database prior to the database lock. These statistical reviews include (1) the comparison of datasets received against the protocol and annotated CRF to ensure that all variables and time points are included; (2) the accountability of every subject at each visit at every site (e.g., which sites have significant dropouts); (3) the primary and secondary analysis variables in terms of unexpected values, outliers, missing values, and other results that may have a significant impact on the analysis; (4) the consistency of use of imperial, metric, or other units and inclusion of unit indicators; (5) other data problems not identified in the data management process;

(6) accurate coding of data format and size; (7) evaluability of subjects based on the various analysis population defined; (8) range check for detecting invalid data that are either below or above expected range; and (9) statistical review using basic SAS procedures. The statistician should identify any data issues that may have an impact on the analysis and submit them to the data manager for resolution. When all identified changes have been made, the statistical review should be repeated based on the revised database for quality assurance.