Lawrence M. Friedman
Curt D. Furberg
David L. DeMets

Fundamentals of Clinical Trials

4th Edition

Springer

Fundamentals of Clinical Trials

Lawrence M. Friedman • Curt D. Furberg
David L. DeMets

Fundamentals of Clinical Trials

Fourth Edition

Springer

Lawrence M. Friedman
Bethesda, MD
USA
l.m.friedman@verizon.net

Curt D. Furberg
Wake Forest University
School of Medicine
Winston-Salem, NC
USA
cfurberg@wfubmc.edu

David L. DeMets
Department of Biostatistics &
Medical Informatics
University of Wisconsin
Madison, WI
USA
demets@biostat.wisc.edu

ISBN 978-1-4419-1585-6 e-ISBN 978-1-4419-1586-3
DOI 10.1007/978-1-4419-1586-3
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: xxxxxxxxxx

© Springer Science+Business Media, LLC 2010
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The clinical trial is "the most definitive tool for evaluation of the applicability of clinical research." It represents "a key research activity with the potential to improve the quality of health care and control costs through careful comparison of alternative treatments" [1]. It has been called on many occasions, "the gold standard" against which all other clinical research is measured.

Although many clinical trials are of high quality, a careful reader of the medical literature will notice that a large number have deficiencies in design, conduct, analysis, presentation, and/or interpretation of results. Improvements have occurred over the past few decades, but too many trials are still conducted without adequate attention to its fundamental principles. Certainly, numerous studies could have been upgraded if the authors had had a better understanding of the fundamentals.

Since the publication of the first edition of this book, a large number of other texts on clinical trials have appeared, most of which are indicated here [2–21]. Several of them, however, discuss only specific issues involved in clinical trials. Additionally, many are no longer current. The purpose of this fourth edition is to update areas in which major progress has been made since the publication of the third edition. We have revised most chapters considerably and added one on ethical issues.

In this book, we hope to assist investigators in improving the quality of clinical trials by discussing fundamental concepts with examples from our experience and the literature. The book is intended both for investigators with some clinical trial experience and for those who plan to conduct a trial for the first time. It is also intended to be used in the teaching of clinical trial methodology and to assist members of the scientific and medical community who wish to evaluate and interpret published reports of trials. Although not a technically oriented book, it may be used as a reference for graduate courses in clinical trials. Those readers who wish to consult more technical books and articles are provided with the relevant literature.

Because of the considerable differences in background and objectives of the intended readership, we have not attempted to provide exercises at the end of each chapter. We have, however, found two exercises to be quite useful and that apply most of the fundamental principles of this text. First, ask students to critique a clinical trial article from the current literature. Second, require students to develop a protocol on a medically relevant research question that is of interest to

the student. These draft protocols often can be turned into protocols that are implemented. This book is also not meant to serve as guide to regulatory requirements. Those differ among countries and frequently change. Rather, as the title indicates, we hope to provide the fundamentals of clinical trials design, conduct, analysis, and reporting.

The first chapter describes the rationale and phases of clinical trials. Chapter 2 is an addition and it covers selected ethical issues. Chapter 3 describes the questions that clinical trials seek to answer and Chap. 4 discusses the populations from which the study samples are derived. The strengths and weaknesses of various kinds of study designs, including noninferiority trials, are reviewed in Chap. 5. The process of randomization is covered in Chap. 6. In Chap. 7, we discuss the importance of and difficulties in maintaining blindness. How the sample size is estimated is covered in Chap. 8. Chapter 9 describes what constitutes the baseline measures. Chapter 10 reviews recruitment techniques and may be of special interest to investigators not having ready access to trial participants. Methods for collecting high quality data and some common problems in data collection are included in Chap. 11. Chapters 12 and 13 focus on the important areas of assessment of adverse events and quality of life. Measures to enhance and monitor participant adherence are presented in Chap. 14. Chapter 15 reviews techniques of survival analysis. Chapter 16 covers data and safety monitoring. Which data should be analyzed? The authors develop this question in Chap. 17 by discussing reasons for not withdrawing participants from analysis. Topics such as subgroup analysis and meta-analysis are also addressed. Chapter 18 deals with phasing out clinical trials, and Chap. 19 with reporting and interpretation of results. Finally, in Chap. 20 we present information about multicenter, including multinational, studies, which have features requiring special attention. Several points covered in the final chapter may also be of value to investigators conducting single center studies.

This book is a collaborative effort and is based on knowledge gained over almost 4 decades in developing, conducting, overseeing, and analyzing data from a number of clinical trials. This experience is chiefly, but not exclusively, in trials of heart and lung diseases, AIDS, and cancer. As a consequence, many of the examples cited are based on work done in these fields. However, the principles are applicable to clinical trials in general. The reader will note that although the book contains examples that are relatively recent, others are quite old. The fundamentals of clinical trials were developed in those older studies, and we cite them because, despite important advances, many of the basic features remain unchanged.

In the first edition, the authors had read or were familiar with much of the relevant literature on the design, conduct, and analysis of clinical trials. Today, that task would be nearly impossible as the literature over the past 3 decades has expanded enormously. The references used in this text are not meant to be exhaustive but rather to include the older literature that established the fundamentals and newer publications that support those fundamentals.

The views expressed in this book are those of the authors and do not necessarily represent the views of the institutions with which the authors have been or are affiliated.

References

1. NIH Inventory of Clinical Trials: Fiscal Year 1979. Volume 1. National Institutes of Health, Division of Research Grants, Research Analysis and Evaluation Branch, Bethesda, MD.
2. Tygstrup N, Lachin J M, Juhl E (eds.). *The Randomized Clinical Trial and Therapeutic Decisions.* New York: Marcel Dekker, 1982.
3. Miké V, Stanley KE (eds.). *Statistics in Medical Research: Methods and Issues, with Applications in Cancer Research.* New York: Wiley, 1982.
4. Pocock SJ. *Clinical Trials – A Practical Approach.* New York: Wiley, 1983.
5. Shapiro SH, Louis TA (eds.). *Clinical Trials – Issues and Approaches.* New York: Marcel Dekker, 1983.
6. Meinert CL. *Clinical Trials: Design, Conduct, and Analysis.* New York: Oxford University Press, 1986.
7. Iber FL, Riley WA, Murray PJ. *Conducting Clinical Trials*. New York: Plenum, 1987.
8. Peace KE (ed.). *Statistical Issues in Drug Research and Development*. New York: Marcel Dekker, 1990.
9. Spilker B. *Guide to Clinical Trials*. New York: Raven, 1991.
10. Spriet A, Dupin-Spriet T, Simon P. *Methodology of Clinical Drug Trials* (2nd edition). Basel: Karger, 1994.
11. Ingelfinger JA, Mosteller F, Thibodeau LA, Ware JH. *Biostatistics in Clinical Medicine* (3rd edition). New York: Macmillan, 1994.
12. Bulpitt CJ. *Randomised Controlled Clinical Trials* (2nd edition). The Hague: Martinus Nijhoff, 1996.
13. Green S, Benedetti J, Crowley J. *Clinical Trials in Oncology* (2nd edition). Boca Raton: CRC, 2002.
14. Chow S-C, Shao J. *Statistics in Drug Research: Methodologies and Recent Developments.* New York: Marcel Dekker, 2002.
15. Geller NL (ed.). *Advances in Clinical Trial Biostatistics.* New York: Marcel Dekker, 2003.
16. Piantadosi S. *Clinical Trials: A Methodologic Perspective* (2nd edition). New York: Wiley, 2005.
17. Matthews JNS. *An Introduction to Randomised Controlled Clinical Trials* (2nd edition). Boca Raton: Chapman & Hall/CRC, 2006.
18. Machin D, Day S, Green S. *Textbook of Clinical Trials* (2nd edition). West Sussex: Wiley, 2006.
19. Hulley SB, Cummings SR, Browner WS, Grady DG. *Designing Clinical Research: An Epidemiologic Approach* (3rd edition). New York: Wolters Kluwer, 2006.
20. Keech A, Gebski V, Pike R (eds.). *Interpreting and Reporting Clinical Trials.* Sidney: Australasian Medical Publishing Company, 2007.
21. Cook TD, DeMets DL (eds.). *Introduction to Statistical Methods for Clinical Trials.* Boca Raton: Chapman & Hall/CRC, Taylor & Francis Group, LLC, 2008.

Acknowledgments

Most of the ideas and concepts discussed in this book represent what we first learned during our years at the National Heart, Lung, and Blood Institute. We are indebted to many colleagues, particularly Dr. William T. Friedewald and the late Dr. Max Halperin, with whom we had numerous hours of discussion for the earlier editions concerning theoretical and operational aspects of the design, conduct, and analysis of clinical trials.

Many have contributed to this edition of the book. We appreciate the efforts of Dr. Michelle Naughton and Dr. Sally Shumaker in revising the chapter on health-related quality of life. Also appreciated are the constructive comments of Drs. Rick Chappell, Mark Espeland, Bengt Furberg, Nancy King, Jeffrey Probstfield, and Dave Reboussin. Finally, we want to particularly acknowledge the outstanding secretarial support of Sarah Hutchens and Sue Parman. Nobody deserves more credit than our wives, Gene Friedman, Birgitta Furberg, and Kathy DeMets. Over the years their support of our professional activities, including the numerous hours put into the four editions of this book, has been unfailing.

Contents

1 Introduction to Clinical Trials 1
Fundamental Point 2
What Is a Clinical Trial? 2
Clinical Trial Phases 3
Phase I Studies 4
Phase II Studies 6
Phase III/IV Trials 7
Why Are Clinical Trials Needed? 8
Problems in the Timing of a Trial 11
Study Protocol 12
References 14

2 Ethical Issues 19
Fundamental Point 20
Planning and Design 20
Does the Question Require a Clinical Trial? 20
Randomization 21
Control Group 22
Protection from Conflict of Interest 23
Informed Consent 24
Conduct 27
Trials in Developing Countries 27
Recruitment 28
Safety and Efficacy Monitoring 29
Early Termination for Other than Scientific or Safety Reasons 29
Privacy and Confidentiality 30
Data Falsification 31
Reporting 31
Publication Bias, Suppression, and Delays 31
Conflicts of Interest and Publication 32
References 32

3 What Is the Question? 37
Fundamental Point 37
Selection of the Questions 38
Primary Question 38
Secondary Questions 38
Adverse Events 39
Ancillary Questions, Substudies 40
Natural History 40
Large, Simple Clinical Trials 41
Superiority vs. Noninferiority Trials 41
Intervention 42
Response Variables 43
Specifying the Question 45
Biomarkers and Surrogate Response Variables 47
General Comments 50
References 51

4 Study Population 55
Fundamental Point 55
Definition of Study Population 55
Rationale 55
Considerations in Defining the Study Population 57
Generalization 62
Recruitment 64
References 65

5 Basic Study Design 67
Fundamental Point 68
Randomized Control Trials 69
Nonrandomized Concurrent Control Studies 72
Historical Controls and Databases 73
Strengths of Historical Control Studies 73
Limitations of Historical Control Studies 74
Role of Historical Controls 78
Cross-Over Designs 79
Withdrawal Studies 81
Factorial Design 82
Group Allocation Designs 83
Hybrid Designs 84
Large, Simple and Pragmatic Clinical Trials 84
Studies of Equivalency and Noninferiority 86
Adaptive Designs 90
References 91

6 The Randomization Process 97
Fundamental Point 97
Fixed Allocation Randomization 98
Simple Randomization 99
Blocked Randomization 100
Stratified Randomization 102
Adaptive Randomization Procedures 105
Baseline Adaptive Randomization Procedures 105
Response Adaptive Randomization 108
Mechanics of Randomization 109
Recommendations 111
Appendix 111
Adaptive Randomization Algorithm 111
References 113

7 Blindness 119
Fundamental Point 119
Types of Trials 119
Unblinded 119
Single-Blind 120
Double-Blind 122
Triple-Blind 123
Protecting the Double-Blind Design 124
Matching of Drugs 125
Coding of Drugs 127
Official Unblinding 127
Inadvertent Unblinding 128
Assessment and Reporting of Blindness 129
References 131

8 Sample Size 133
Fundamental Point 133
Statistical Concepts 134
Dichotomous Response Variables 139
Two Independent Samples 139
Paired Dichotomous Response 144
Adjusting Sample Size to Compensate for Nonadherence 145
Sample Size Calculations for Continuous Response Variables 147
Two Independent Samples 147
Paired Data 148
Sample Size for Repeated Measures 150
Sample Size Calculations for "Time to Failure" 152
Sample Size for Testing "Equivalency" or Noninferiority of Interventions 155

Sample Size for Cluster Randomization 157
Estimating Sample Size Parameters 159
Multiple Response Variables 161
References 162

9 Baseline Assessment 169
Fundamental Point 169
Uses of Baseline Data 169
Baseline Comparability 170
Stratification 171
Subgrouping 171
Pharmacogenetics 172
Changes of Baseline Measurement 173
Natural History Analyses 173
What Constitutes a True Baseline Measurement? 174
Screening for Participants 174
Regression Toward the Mean 175
Interim Events 176
Uncertainty About Qualifying Diagnosis 177
Contamination of the Intervention 178
Assessment of Baseline Comparability 179
Testing for Baseline Imbalance 180
References 181

10 Recruitment of Study Participants 183
Fundamental Point 183
Considerations Before Participant Enrollment 184
Selection of Study Sample 184
Common Recruitment Problems 184
Planning 186
Recruitment Sources 188
Conduct 190
Monitoring 192
Approaches to Lagging Recruitment 194
References 197

11 Data Collection and Quality Control 199
Fundamental Point 200
Problems in Data Collection 201
Major Types 201
Minimizing Poor Quality Data 203
Design of Protocol and Manual 203
Development of Forms 203
Training and Certification 204
Pretesting 205

Techniques to Reduce Variability 206
Data Entry 206
Quality Monitoring 207
Monitoring of Forms 208
Monitoring of Procedures 208
Monitoring of Drug Handling 209
Audits Leader 210
References 212

12 Assessing and Reporting Adverse Events 215
Fundamental Point 216
Clinical Trials in the Assessment of Adverse Events 216
Strengths 216
Limitations in Identification of SAEs 217
Determinants of Adverse Events 218
Definitions 218
Classification of Adverse Events 219
Ascertainment 220
Dimensions 221
Length of Follow-Up 221
Analyzing Adverse Events 223
Types of Analysis 223
Analysis of Data from Nonadherent Participants 224
Reporting of Adverse Events 224
Scientific 224
Published Reports 225
Regulatory 226
Identification of SAEs 227
Potential Solutions 228
References 229

13 Assessment of Health-Related Quality of Life 233
Fundamental Point 234
Defining Health-Related Quality of Life 234
Primary HRQL Dimensions 235
Additional HRQL Dimensions 236
Uses of Health-Related Quality of Life 237
Methodological Issues 239
Trial Design 239
Study Population 240
Intervention 240
Selection of HRQL Instruments 241
Modes of Administration 242
Frequency of Assessment (Acute Vs. Chronic) 243
Symptom Expression (Episodic Vs. Constant) 244

Functional Impact (Present Vs. Absent) 244
Interpretation 244
Scoring of HRQL Measures 245
Determining the Significance of HRQL Measures 245
Utility Measures/Preference Scaling 246
References 247

14 Participant Adherence 251
Fundamental Point 252
Considerations Before Participant Enrollment 253
Design Factors 253
Participant Factors 255
Maintaining Good Participant Adherence 258
Adherence Monitoring 262
Dealing with Low Adherence 265
Special Populations 266
References 267

15 Survival Analysis 269
Fundamental Point 269
Estimation of the Survival Curve 270
Kaplan–Meier Estimate 274
Cutler–Ederer Estimate 278
Comparison of Two Survival Curves 279
Point-by-Point Comparison 279
Comparison of Median Survival Times 279
Total Curve Comparison 280
Generalizations 285
Covariate Adjusted Analysis 287
References 290

16 Monitoring Response Variables 293
Fundamental Point 295
Monitoring Committee 295
Repeated Testing for Significance 299
Decision for Early Termination 301
Decision to Extend a Trial 310
Statistical Methods Used in Monitoring 313
Classical Sequential Methods 314
Group Sequential Methods 315
Flexible Group Sequential Procedures: Alpha Spending Functions 318
Applications of Group Sequential Boundaries 321

Asymmetric Boundaries 324
Curtailed Sampling and Conditional Power Procedures 325
Other Approaches 330
Trend Adaptive Designs and Sample Size Adjustments 332
References 334

17 Issues in Data Analysis 345
Fundamental Point 345
Which Participants Should Be Analyzed? 346
Ineligibility 347
Nonadherence 350
Missing or Poor Quality Data 355
Competing Events 362
Composite Outcomes 363
Covariate Adjustment 364
Surrogates as Covariates 365
Baseline Variables as Covariates 368
Subgroup Analyses 371
Not Counting Some Events 376
Comparison of Multiple Variables 377
Use of Cutpoints 378
Noninferiority Trial Analysis 379
Meta-Analysis of Multiple Studies 382
Rationale and Issues 382
Statistical Methods 386
Analysis Following Trend Adaptive Designs 389
Appendix 389
Mantel–Haenszel Statistic 389
References 390

18 Closeout 399
Fundamental Point 399
Termination Procedures 399
Planning 399
Scheduling of Closeout Visits 400
Final Response Ascertainment 401
Transfer of Posttrial Care 402
Data and Other Study Material 403
Cleanup and Verification 403
Storage 404
Dissemination of Results 405
Poststudy Follow-Up 407
References 409

19 Reporting and Interpreting of Results 411
Fundamental Point 412
Guidelines for Reporting 413
Authorship 413
Disclosure of Conflict of Interest 414
Presentation of Data 414
Interpretation 415
Publication Bias 416
Did the Trial Work as Planned? 417
Baseline Comparability 417
Blindness 417
Adherence and Concomitant Treatment 418
Limitations 418
Analysis 419
How Do the Findings Compare with Results from Other Studies? 420
What are the Clinical Implications of the Findings? 421
References 422

20 Multicenter Trials 427
Fundamental Point 427
Reasons for Multicenter Trials 428
Conduct of Multicenter Trials 429
Globalization of Trials 436
General Comments 437
References 438

Index 443

Chapter 1
Introduction to Clinical Trials

The evolution of the modern clinical trial dates back to the eighteenth century [1, 2]. Lind, in his classical study on board the *Salisbury*, evaluated six treatments for scurvy in 12 patients. One of the two who was given oranges and lemons recovered quickly and was fit for duty after 6 days. The second was the best recovered of the others and was assigned the role of nurse to the remaining ten patients. Several other comparative studies were also conducted in the eighteenth and nineteenth centuries. The comparison groups comprised literature controls, other historical controls, and concurrent controls [2].

The concept of randomization was introduced by Fisher and applied in agricultural research in 1926 [3]. The first clinical trial that used a form of random assignment of participants to study groups was reported in 1931 by Amberson et al. [4]. After careful matching of 24 patients with pulmonary tuberculosis into comparable groups of 12 each, a flip of a coin determined which group received sanocrysin, a gold compound commonly used at that time. The British Medical Research Council trial of streptomycin in patients with tuberculosis, reported in 1948, was the first to use random numbers in the allocation of individual participants to experimental and control groups [5, 6].

The principle of blindness was also introduced in the trial by Amberson et al. [4]. The participants were not aware of whether they received intravenous injections of sanocrysin or distilled water. In a trial of cold vaccines in 1938, Diehl and coworkers [7] referred to the saline solution given to the subjects in the control group as a placebo.

One of the early trials from the National Cancer Institute of the National Institutes of Health in 1960 randomly assigned patients with leukemia to either 6-azauracil or placebo. No treatment benefit was observed in this double-blind trial [8].

It is only in the past several decades that the clinical trial has emerged as the preferred method in the evaluation of medical interventions. Techniques of implementation and special methods of analysis have been developed during this period. Many of the principles have their origins in work by Hill [9–12]. For a brief history of key development in clinical trials, see Chalmers [13].

The authors of this book have spent their careers at the U.S. National Institutes of Health and/or academia. Therefore, many of the examples reflect this experience.

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_1, © Springer Science+Business Media, LLC 2010

We also cite papers which review the history of clinical trials development at the NIH [14–18].

The purpose of this chapter is to define clinical trials, review the need for them, and discuss timing and phasing of clinical trials.

Fundamental Point

A properly planned and executed clinical trial is a powerful experimental technique for assessing the effectiveness of an intervention.

What Is a Clinical Trial?

We define a clinical trial as a *prospective study comparing the effect and value of intervention(s) against a control in human beings.* Note that a clinical trial is *prospective*, rather than retrospective. Study participants must be followed forward in time. They need not all be followed from an identical calendar date. In fact, this will occur only rarely. Each participant however, must be followed from a well-defined point in time, which becomes time zero or baseline for the study. This contrasts with a case-control study, a type of retrospective observational study in which participants are selected on the basis of presence or absence of an event or condition of interest. By definition, such a study is not a clinical trial. People can also be identified from hospital records or other data sources, and subsequent records can be assessed for evidence of new events. This is not considered to be a clinical trial since the participants are not directly observed from the moment of initiation of the study and at least some of the follow-up data are retrospective.

A clinical trial must employ one or more *intervention* techniques. These may be single or combinations of diagnostic, preventive, or therapeutic drugs, biologics, devices, regimens, or procedures. Intervention techniques should be applied to participants in a standard fashion in an effort to change some aspect. Follow-up of people over a period of time without active intervention may measure the natural history of a disease process, but it does not constitute a clinical trial. Without active intervention the study is observational because no experiment is being performed.

Early phase studies may be controlled or uncontrolled. Although common terminology refers to phase I and phase II trials, because they are sometimes uncontrolled, we will refer to them as clinical studies. A trial, using our definition, contains a *control* group against which the intervention group is compared. At baseline, the control group must be sufficiently similar in relevant respects to the intervention group in order that differences in outcome may reasonably be attributed to the action of the intervention. Methods for obtaining an appropriate control group are discussed in a later chapter. Most often a new intervention is compared with, or used along with, best current standard therapy. Only if no such standard exists or, for several reasons discussed in Chap. 2, is not available, is it appropriate for the participants in the intervention

group to be compared to participants who are on no active treatment. "No active treatment" means that the participant may receive either a placebo or no treatment at all. Obviously, participants in all groups may be on a variety of additional therapies and regimens, so-called concomitant treatments, which may be either self-administered or prescribed by others (e.g., personal physicians).

For purposes of this book, only studies on *human beings* will be considered as clinical trials. Certainly, animals (or plants) may be studied using similar techniques. However, this book focuses on trials in people, and each clinical trial must therefore incorporate participant safety considerations into its basic design. Equally important is the need for, and responsibility of, the investigator to fully inform potential participants about the trial, including information about potential benefits, harms, and treatment alternatives [19–22]. See Chap. 2 for further discussion of ethical issues.

Unlike animal studies, in clinical trials the investigator cannot dictate what an individual should do. He can only strongly encourage participants to avoid certain medications or procedures which might interfere with the trial. Since it may be impossible to have "pure" intervention and control groups, an investigator may not be able to compare interventions, but only intervention strategies. Strategies refer to attempts at getting all participants to adhere, to the best of their ability, to their originally assigned intervention. When planning a trial, the investigator should recognize the difficulties inherent in studies with human subjects and attempt to estimate the magnitude of participants' failure to adhere strictly to the protocol. The implications of less than perfect adherence are considered in Chap. 8.

As discussed in Chaps. 6 and 7, *the ideal clinical trial is one that is randomized and double-blind.* Deviation from this standard has potential drawbacks, which will be discussed in the relevant chapters. In some clinical trials, compromise is unavoidable, but often deficiencies can be prevented by adhering to fundamental features of design, conduct, and analysis.

A number of people distinguish between demonstrating "efficacy" of an intervention and "effectiveness" of an intervention. They also refer to "explanatory" trials, as opposed to "pragmatic" or "practical" trials. Efficacy or explanatory trials refer to what the intervention accomplishes in an ideal setting. The term is sometimes used to justify not using an "intention-to-treat" analysis. As discussed in Chaps. 8 and 17, that is insufficient justification. Effectiveness or pragmatic trials refer to what the intervention accomplishes in actual practice, taking into account incomplete adherence to the protocol. We do not consider this distinction between trials as important as the proper design, conduct, and analysis of all trials in order to answer important clinical or public health questions, regardless of the setting in which they are done.

Clinical Trial Phases

While we focus on the design and analysis of randomized trials comparing the effectiveness of two or more interventions, several steps or phases of clinical research must occur before this comparison can be implemented. Classically, trials of pharmaceutical

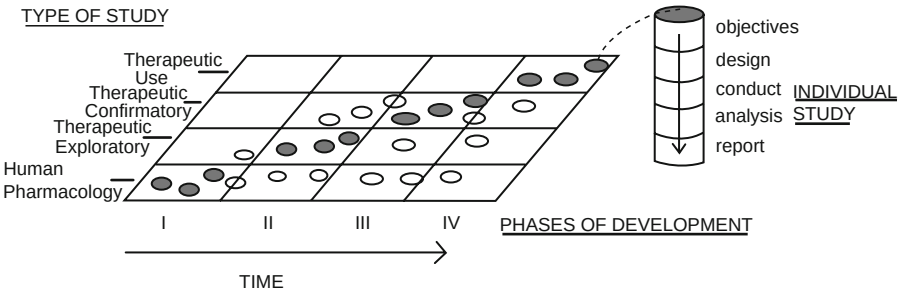


Fig. 1.1 Correlation between development phases and types of study [23]

agents have been divided into phases I–IV. Studies with other kinds of interventions, particularly those involving behavior or lifestyle change or surgical approaches, will often not fit neatly into those phases. In addition, even trials of drugs may not fit into a single phase. For example, some may blend from phase I to phase II or from phase II to phase III. Therefore, it may be easier to think of early phase studies and late phase studies. Nevertheless, because they are still in common use, and because early phase studies, even if uncontrolled, may provide information essential for the conduct of late phase trials, the phases are defined below.

An excellent summary of phases of clinical trials and the kinds of questions addressed at each phase was prepared by the International Conference on Harmonisation [23]. Figure 1.1, taken from that document, illustrates that research goals can overlap with more than one study phase.

Thus, although human pharmacology studies, which examine drug tolerance, metabolism, and interactions, and describe pharmacokinetics and pharmacodynamics, are generally done as phase I, some pharmacology studies may be done in other trial phases. Therapeutic exploratory studies, which look at the effects of various doses and typically use biomarkers as the outcome, are generally thought of as phase II. However, sometimes, they may be incorporated into other phases. The usual phase III trial consists of therapeutic confirmatory studies, which demonstrate clinical use and establish the safety profile. But such studies may also be done in phase II or phase IV trials. Therapeutic use studies, which examine the drug in broad or special populations and seek to identify uncommon adverse events, are almost always phase IV trials.

Phase I Studies

Although useful preclinical information may be obtained from in vitro studies or animal models, early data must be obtained in humans. People who participate in phase I studies generally are healthy volunteers but may also be patients who have typically already tried and failed to improve on the existing standard therapies.

Phase I studies attempt to estimate tolerability and characterize pharmacokinetics and pharmacodynamics. They focus on questions such as bioavailability and body compartment distribution. They also provide preliminary assessment of drug activity [23]. Buoen et al. reviewed 105 phase I dose-escalation studies in several medical disciplines that used healthy volunteers [24]. Despite the development of new designs, primarily in the field of cancer research, most of the studies in the survey employed simple dose-escalation approaches.

Often, one of the first steps in evaluating drugs is to estimate how large a dose can be given before unacceptable toxicity is experienced by participants [25–30]. This dose is usually referred to as the maximally tolerated dose. Much of the early literature has discussed how to extrapolate animal model data to the starting dose in humans [31] or how to step up the dose levels to achieve the maximally tolerated dose.

In estimating the maximally tolerated dose, the investigator usually starts with a very low dose and escalates the dose until a prespecified level of toxicity is obtained. Typically, a small number of participants, usually three, are entered sequentially at a particular dose. If no specified level of toxicity is observed, the next predefined higher dose level is used. If unacceptable toxicity is observed in any of the three participants, an additional number of participants, usually three, are treated at the same dose. If no further toxicity is seen, the dose is escalated to the next higher dose. If an additional unacceptable toxicity is observed, then the dose escalation is terminated and that dose, or perhaps the previous dose, is declared to be the maximally tolerated dose. This particular design assumes that the maximally tolerated dose occurs when approximately one-third of the participants experience unacceptable toxicity. Variations of this design exist, but most are similar.

Some [29, 32–34] have proposed more sophisticated designs in cancer research that specify a sampling scheme for dose escalation and a statistical model for the estimate of the maximally tolerated dose and its standard error. The sampling scheme must be conservative in dose escalation so as not to overshoot the maximally tolerated dose by very much, but at the same time be efficient in the number of participants studied. Many of the proposed schemes utilize a step-up/step-down approach; the simplest being an extension of the previously mentioned design to allow step-downs instead of termination after unacceptable toxicity, with the possibly of subsequent step-ups. Further increase or decrease in the dose level depends on whether or not toxicity is observed at a given dose. Dose escalation stops when the process seems to have converged around a particular dose level. Once the data are generated, a dose response model is fit to the data and estimates of the maximally tolerated dose can be obtained as a function of the specified probability of a toxic response [29].

Bayesian approaches have also been developed [35, 36]. These involve methods employing continual reassessment [32, 37] and escalation with overdose control [38]. Bayesian methods involve the specification of the investigators' prior opinion about the agent's dose-toxicity profile, which is then used to select starting doses and escalation rules. The most common Bayesian phase I design is called the continual reassessment method, [32] in which the starting dose is set to the prior estimate

of the maximally tolerated dose. After the first cohort of participants (typically of size 1, 2, or 3, though other numbers are possible), the estimate is updated and the next participant(s) assigned to that estimate. The process is repeated until a prespecified number of participants have been assigned. The dose at which a hypothetical additional participant would be assigned constitutes the final estimate of the maximally tolerated dose. Bayesian methods that constrain the number of total toxicities have also been developed (escalation with overdose control) as have designs that allow for two or more treatments [39] and as have methods that allow for incomplete follow-up of long-term toxicities (time-to-event continual reassessment method) (40). Many variations have been proposed. An advantage of Bayesian phase I designs is that they are very flexible, allowing risk factors and other sources of information to be incorporated into escalation decisions. A disadvantage is their complexity, leading to unintuitive dose assignment rules.

A detailed description of the design and conduct of dose escalating trials for treatments of cancer is found in Chaps. 1–5 of a book edited by Crowley and Ankerst [41]. A book edited by Ting contains a more general discussion of dose-selection approaches [42].

Phase II Studies

Once a dose or range of doses is determined, the next goal is to evaluate whether the drug has any biological activity or effect. The comparison may consist of a concurrent control group, historical controls, or pretreatment status versus post-treatment status. Because of uncertainty with regard to dose–response, phase II studies may also employ several doses, with perhaps four or five intervention arms. They will look, for example, at the relationship between blood level and activity. Genetic testing is common, particularly when there is evidence of variation in metabolic rate. Participants in phase II studies are usually carefully selected, with narrow inclusion criteria [23].

The phase II design depends on the quality and adequacy of the phase I study. The results of the phase II study will, in turn, be used to design the comparative phase III trial. The statistical literature for phase II studies, which had been rather limited [43–49] has expanded [50, 51] and, as with phase I studies, includes Bayesian methods [52, 53].

One of the traditional phase II designs in cancer is based on the work of Gehan, [43] which is a version of a two stage design. In the first stage, the investigator attempts to rule out drugs which have no or little biologic activity. For example, he may specify that a drug must have some minimal level of activity, say, in 20% of participants. If the estimated activity level is less than 20%, he chooses not to consider this drug further, at least not at that maximally tolerated dose. If the estimated activity level exceeds 20%, he will add more participants to get a better estimate of the response rate. A typical study for ruling out a 20% or lower response rate enters 14 participants. If no response is observed in the first 14 participants, the drug is

considered not likely to have a 20% or higher activity level. The number of additional participants added depends on the degree of precision desired, but ranges from 10 to 20. Thus, a typical cancer phase II study might include fewer than 30 people to estimate the response rate. As is discussed in Chap. 8, the precision of the estimated response rate is important in the design of the controlled trial. In general, phase II studies are smaller than they ought to be.

Some [29, 44, 54] have proposed designs which have more stages or a sequential aspect. Others [47, 55] have considered hybrids of phase II and phase III designs in order to enhance efficiency. While these designs have desirable statistical properties, the most vulnerable aspect of phase II, as well as phase I studies, is the type of person enrolled. Usually, phase II studies have more exclusion criteria than phase III comparative trials. Furthermore, the outcome in the phase II study (e.g., tumor response) may be different than that used in the definitive comparative trial (e.g., survival). Refinements may include time to failure [51] and unequal numbers of participants in the various stages of the phase II study [56]. Bayesian designs for phase II trials require prior estimates, as was the case for phase I studies, but differ in that they are priors of efficacy measures for the dose or doses to be investigated rather than of toxicity rates. Priors are useful for incorporating historical data into the design and analysis of phase II trials. Methods are available for continuous [57], bivariate [57], and survival outcomes [58]. These methods can account not only for random variations in participant responses within institutions but also for systematic differences in outcomes between institutions in multicenter trials or when several control groups are combined. They also acknowledge the fact that historical efficacy measures of the control are estimated with error. This induces larger sample sizes than in trials which assume efficacy of the control to be known, but with correspondingly greater resistance to false positive and false negative errors. Bayesian methods can also be used in a decision-theoretic fashion to minimize a prespecified combination of these errors for a given sample size [59, 60].

Phase III/IV Trials

The phase III trial is the clinical trial defined earlier in the chapter. It is generally designed to assess the effectiveness of the new intervention and thereby, its value in clinical practice. The focus of most of this book is on phase III and other late phase trials. However, many design assumptions depend on information obtained from phase I and phase II studies, or some combination of early phase studies.

Phase III trials of chronic conditions or diseases often have a short follow-up period for evaluation, relative to the period of time the intervention might be used in practice. In addition, they focus on effectiveness, but knowledge of safety is also necessary to evaluate fully the proper role of an intervention. A procedure or device may fail after a few years and have adverse sequelae for the patient. Thus, long-term surveillance of an intervention believed to be effective in phase III trials is necessary. Such long-term studies or studies conducted after regulatory agency approval

of the drug or device, are referred to as phase IV trials. Drugs may be approved on the basis of intermediate outcomes or biomarkers, such as blood pressure or cholesterol lowering. They may also be approved after relatively short-term studies (weeks or months), even though in practice, in the case of chronic conditions, they may be taken for years or even decades. Even late phase clinical trials are limited in size to several hundreds or thousands (at most, a few tens of thousands) of participants. Yet the approved drugs or devices will possibly be used by millions of people. This combination of incomplete information about clinical outcomes, relatively short duration, and limited size means that sometimes the balance between benefit and harm becomes clear only when larger phase IV studies are done, or when there is greater clinical experience. One example is some of the cyclooxygenase 2 (COX 2) inhibitors, which had been approved for arthritis pain, but only disclosed cardiovascular problems after larger trials were done. These larger trials were examining the effects of the COX 2 inhibitors on prevention of colon cancer in those with polyps [61, 62]. Similarly, only after they had been on the market were thiazolidinediones, a class of drugs used for diabetes, found to be associated with an increase in heart failure [63].

Regulatory agency approval of drugs, devices, and biologics may differ because, at least in the United States, the regulations for these different kinds of interventions are based on different laws. For example, FDA approval of drugs depends greatly on at least one well-designed clinical trial plus supporting evidence (often, another clinical trial). Approval of devices relies less on clinical trial data and more on engineering characteristics of the device, including similarity with previously approved devices. Devices, however, are often implanted, and unless explanted, may be present for the life of the participant. Therefore, there are urgent needs for truly long-term data on performance of devices in vivo. Assessment of devices also depends, more so than drugs, on the skill of the person performing the implantation. As a result, the results obtained in a clinical trial, which typically uses only well-trained investigators, may not provide an accurate balance of harm and benefit in actual practice.

The same caution applies to clinical trials of procedures of other sorts, whether surgical or lifestyle intervention, where only highly skilled practitioners are investigators. But unlike devices, procedures may have little or no regulatory oversight although those paying for care often consider the evidence.

Why Are Clinical Trials Needed?

A clinical trial is the most definitive method of determining whether an intervention has the postulated effect. Only seldom is a disease or condition so completely characterized that people fully understand its natural history and can say, from knowledge of pertinent variables, what the subsequent course of a group of patients will be. Even more rarely can a clinician predict with certainty the outcome in individual patients. By outcome is meant not simply that an individual will die, but when,

and under what circumstances; not simply that he will recover from a disease, but what complications of that disease he will suffer; not simply that some biological variable has changed, but to what extent the change has occurred. Given the uncertain knowledge about disease course and the usual large variations in biological measures, it is often difficult to say on the basis of uncontrolled clinical observation whether a new treatment has made a difference to outcome and, if it has, what the magnitude is. A clinical trial offers the possibility of such judgment because there exists a control group – which, ideally, is comparable to the intervention group in every way except for the intervention being studied.

The consequences of not conducting appropriate clinical trials at the proper time can be serious or costly. An example was the uncertainty as to the efficacy and safety of digitalis in congestive heart failure. Only in the 1990s, after the drug had been used for over 200 years, was a large clinical trial evaluating the effect of digitalis on mortality mounted [64]. Intermittent positive pressure breathing became an established therapy for chronic obstructive pulmonary disease without good evidence of benefits. One trial suggested no major benefit from this very expensive procedure [65]. Similarly, high concentration of oxygen was used for therapy in premature infants until a clinical trial demonstrated its harm [66]. A clinical trial can determine the incidence of adverse effects or complications of the intervention. Few interventions, if any, are entirely free of undesirable effects. However, drug toxicity might go unnoticed without the systematic follow-up measurements obtained in a clinical trial of sufficient size. The Cardiac Arrhythmia Suppression Trial documented that commonly used antiarrhythmic drugs were harmful in patients who had a history of myocardial infarction, and raised questions about routine use of an entire class of antiarrhythmic agents [67]. Corticosteroids had been commonly used to treat people with traumatic brain injury. Small clinical trials were inconclusive, and a meta-analysis of 16 trials showed no difference in mortality between corticosteroids and control [68]. Because of the uncertainty as to benefit, a large clinical trial was conducted. This trial, with far more participants than the others combined, demonstrated a significant 18% relative increase in mortality at 14 days [69] and a 15% increase at 6 months [70]. As a result, an update of the meta-analysis recommended against the routine use of corticosteroids in people with head injury [71].

In the final evaluation, an investigator must compare the benefit of an intervention with its other, possibly unwanted effects in order to decide whether, and under what circumstances, its use should be recommended. The cost implications of an intervention, particularly if there is limited benefit, must also be considered. Several studies have indicated that drug eluting stents have somewhat less restenosis than bare metal stents in percutaneous coronary intervention [72, 73]. The cost difference, however, can be considerable, especially since more than one stent is typically inserted. Are the added benefits, which may be defined and measured in different ways, of the most expensive interventions worth the extra cost? Such assessments are not statistical in nature. They must rely on the judgment of the investigator and the medical practitioner as well as on those who pay for medical care. Clinical trials rarely fully assess costs of the interventions and associated

patient care, which change over time, and cannot make these decisions; they can only provide data so that decisions are evidence-based.

Those suffering from or treating life-threatening diseases for which there are no known effective therapies often argue that controlled clinical trials are not needed and that they have a right to experimental interventions. Because there may be little hope of cure or even improvement, patients and their physicians want to have access to new interventions, even if those interventions have not been shown to be safe and effective by means of the usual clinical trial. They want to be in studies of these interventions, with the expectation that they will receive the new treatment, rather than the control (if there is a control group). Those with the acquired immunodeficiency syndrome (AIDS) used to make the case forcefully that traditional clinical trials are not the sole legitimate way of determining whether interventions are useful [74–77]. This is undeniably true, and clinical trial researchers need to be willing to modify, when necessary, aspects of study design or management. Many have been vocal in their demands that once a drug or biologic has undergone some minimal investigation, it should be available to those with life-threatening conditions, should they desire it, even without late phase clinical trial evidence [78]. If the patient community is unwilling to participate in clinical trials conducted along traditional lines, or in ways that are scientifically "pure," trials are not feasible and no information will be forthcoming. Investigators need to involve the relevant communities or populations at risk, even though this could lead to some compromises in design and scientific purity. Investigators need to decide when such compromises so invalidate the results that the study is not worth conducting. It should be noted that the rapidity with which trial results are demanded, the extent of community involvement, and the consequent effect on study design, can change as knowledge of the disease increases, as at least partially effective therapy becomes available, and as understanding of the need for valid research designs, including clinical trials, develops. This happened to a great extent with AIDS trials.

Although investigators should design clinical trials using the fundamentals discussed in this book, they must consider the context in which the trial is being conducted. The nature of the disease or condition being studied and the population and setting in which it is being done will influence the outcomes that are assessed, the kind of control, the size, the duration, and many other factors.

Clinical trials are conducted because it is expected that they will influence practice. The literature on this is limited [79–85], and it is unclear how much of the reduction in mortality or morbidity due to better preventive and treatment approaches can be directly attributed to the results of clinical trials [86]. For example, the decline in stroke mortality in the U.S. and elsewhere began before there was effective or widespread treatment of hypertension and well before clinical trials demonstrated the benefits of antihypertensive agents [87]. It is undoubtedly true that multiple reasons exist, and it is not possible to clearly define the societal importance of clinical trials. Further, the influence of trials depends on direction of the findings, means of dissemination of the results, existence of evidence from other relevant research, and probably other factors. However, well-designed clinical trials can certainly have pronounced effects on clinical practice [80].

There is no such thing as a perfect study. However, a well thought-out, well-designed, appropriately conducted and analyzed clinical trial is an effective tool. While even well-designed clinical trials are not infallible, they can provide a sounder rationale for intervention than is obtainable by other research methods. On the other hand, poorly designed and conducted trials can be misleading. Also, without supporting evidence, no single study ought to be definitive. When interpreting the results of a trial, consistency with data from laboratory, animal, epidemiological, and other clinical research must be considered.

Some have claimed that observational studies provide the "correct" answer more often than not and that therefore clinical trials are often superfluous [88, 89]. Others have pointed out that sometimes, results of observational studies and clinical trials are inconsistent. Observational studies, many of them large, suggested that use of antioxidants would reduce the risk of cancer and heart disease. They began to be widely used as a result. Later, large randomized controlled trials evaluating many of the antioxidants demonstrated no benefit or even harm [90]. Similarly, because of the results from observational studies, hormone therapy was advocated for postmenopausal women as a way to prevent or reduce heart disease. Results of large clinical trials [91–93] cast considerable doubt on the findings from the observational studies. Whether the differences are due to the inherent limitations of observational studies (see Chap. 5), to limitations in the designs of the clinical trials, or some combination has been debated. Regardless, anyone considering taking (or administering) antioxidants for the purpose of heart disease or cancer prevention, or hormone replacement therapy to prevent heart disease, must carefully evaluate the results of the trials.

We believe that pitting one kind of clinical research against another is inappropriate. Both observational epidemiology studies and clinical trials have their strengths and weaknesses; both have their place [94]. Proper understanding of the strengths and weaknesses of clinical trials, and how the results of well-designed and conducted trials can be used in conjunction with other research methodologies, is by far the best way of improving public health and scientific understanding.

Problems in the Timing of a Trial

Once drugs and procedures of unproved clinical benefit have become part of general medical practice, performing an adequate clinical trial becomes difficult ethically and logistically. Some people advocated instituting clinical trials as early as possible in the evaluation of new therapies [95, 96]. The trials, however, must be feasible. Assessing feasibility takes into account several factors. Before conducting a trial, an investigator needs to have the necessary knowledge and tools. He must know something about the safety of the intervention and what outcomes to assess and have the techniques to do so. Well-run clinical trials of adequate magnitude are costly and should be done only when preliminary evidence of the efficacy of an intervention looks promising enough to warrant the effort and expense involved.

Another aspect of timing is consideration of the relative stability of the intervention. If active research will be likely to make the intended intervention outmoded in a short time, studying such an intervention may be inappropriate. This is particularly true in long-term clinical trials, or studies that take many months to develop. One of the criticisms of trials of surgical interventions has been that surgical methods are constantly being improved. Evaluating an operative technique of several years past, when a study was initiated, may not reflect the current status of surgery [97–99].

These issues were raised in connection with the Veterans Administration study of coronary artery bypass surgery [100]. The trial showed that surgery was beneficial in subgroups of patients with left main coronary artery disease and three vessel disease, but not overall [100–102]. Critics of the trial argued that when the trial was started, the surgical techniques were still evolving. Therefore, surgical mortality in the study did not reflect what occurred in actual practice at the end of the long-term trial. In addition, there were wide differences in surgical mortality between the cooperating clinics [103] that may have been related to the experience of the surgeons. Defenders of the study maintained that the surgical mortality in the Veterans Administration hospitals was not very different from the national experience at the time (104). In the Coronary Artery Surgery Study [105], surgical mortality was lower than in the Veterans Administration trial, reflecting better technique. The control group mortality, however, was also lower.

Review articles show that surgical trials have been successfully undertaken [106, 107] and, despite challenges, can and should be conducted [108, 109]. While the best approach might be to postpone a trial until a procedure has reached a plateau and is unlikely to change greatly, such a postponement will probably mean waiting until the procedure has been widely accepted as efficacious for some indication, thus making it difficult, if not impossible to conduct the trial. However, as noted by Chalmers and Sacks, [110] allowing for improvements in operative techniques in a clinical trial is possible. As in all aspects of conducting a clinical trial, judgment must be used in determining the proper time to evaluate an intervention.

Study Protocol

Every well-designed clinical trial requires a protocol. The study protocol can be viewed as a written agreement between the investigator, the participant, and the scientific community. The contents provide the background, specify the objectives, and describe the design and organization of the trial. Every detail explaining how the trial is carried out does not need to be included, provided that a comprehensive manual of procedures contains such information. The protocol serves as a document to assist communication among those working in the trial. It should also be made available to others upon request.

The protocol should be developed before the beginning of participant enrollment and should remain essentially unchanged except perhaps for minor updates.

Careful thought and justification should go into any changes. Major revisions which alter the direction of the trial should be rare. If they occur, the rationale behind such changes needs to be clearly described. An example is the Cardiac Arrhythmia Suppression Trial, which, on the basis of important study findings, changed intervention, participant eligibility criteria, and sample size [111].

Registration of all late phase trials and many early phase studies is now advocated, and indeed required by many journals and sponsors. Journals will not publish results of trials or study design papers unless the study has been registered at one of many sites, such as ClinicalTrials.gov [112] and the WHO International Clinical Trials Registry Platform (ICTRP) [113]. The U.S. National Institutes of Health requires that trials it funds be registered [114], as does the Food and Drug Administration for trials it oversees [115]. The registry sites have, at a minimum, information about the study population, intervention and control, response variables, and other key elements of the study design. See Chap. 18 for a further discussion of trial registration. We applaud the practice of registration, and encourage all investigators to go further by including links to their protocols at the registry sites.

Topic headings of a typical protocol, which also serve as an outline of the subsequent chapters in this book, are given below:

A. Background of the study
B. Objectives
    1. Primary question and response variable
    2. Secondary questions and response variables
    3. Subgroup hypotheses
    4. Adverse effects
C. Design of the study
    1. Study population
        a. Inclusion criteria
        b. Exclusion criteria
    2. Sample size assumptions and estimates
    3. Enrollment of participants
        a. Informed consent
        b. Assessment of eligibility
        c. Baseline examination
        d. Intervention allocation (e.g., randomization method)
    4. Intervention(s)
        a. Description and schedule
        b. Measures of compliance
    5. Follow-up visit description and schedule
    6. Ascertainment of response variables
        a. Training
        b. Data collection
        c. Quality control
    7. Safety Assessment
        a. Type and frequency

   b. Instruments
   c. Reporting
8. Data analysis
   a. Interim monitoring
   b. Final analysis
9. Termination policy

D. Organization
1. Participating investigators
   a. Statistical unit or data coordinating center
   b. Laboratories and other special units
   c. Clinical center(s)
2. Study administration
   a. Steering committees and subcommittees
   b. Data monitoring committee
   c. Funding organization

Appendices
Definitions of eligibility criteria
Definitions of response variables
Informed Consent Form

References

1. Bull JP. The historical development of clinical therapeutic trials. *J Chronic Dis* 1959;10:218–248.
2. Lilienfeld AM. Ceteris paribus: the evolution of the clinical trial. *Bull Hist Med* 1982;56:1–18.
3. Box JF. R. A. Fisher and the design of experiments, 1922–1926. *Am Stat* 1980;34:1–7.
4. Amberson JB, Jr, McMahon BT, Pinner M. A clinical trial of sanocrysin in pulmonary tuberculosis. *Am Rev Tuberc* 1931;24:401–435.
5. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *Br Med J* 1948;2:769–782.
6. Hart PD. Letter to the editor: Randomised controlled clinical trials. *Br Med J* 1991;302:1271–1272.
7. Diehl HS, Baker AB, Cowan DW. Cold vaccines; an evaluation based on a controlled study. *JAMA* 1938;111:1168–1173.
8. Freireich EJ, Frei E, III, Holland JF, et al. Evaluation of a new chemotherapeutic agent in patients with “advanced refractory” acute leukemia: studies of 6-azauracil. *Blood* 1960;16:1268–1278.
9. Hill AB. The clinical trial. *Br Med Bull* 1951;7:278–282.
10. Hill AB. The clinical trial. *N Engl J Med* 1952;247:113–119.
11. Hill AB. *Statistical Methods of Clinical and Preventive Medicine*. 1962; Oxford University Press, New York.
12. Doll R. Clinical trials: retrospect and prospect. *Stat Med* 1982;1:337–344.
13. Chalmers I. Comparing like with like: some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments. *Int J Epidemiol* 2001;30:1156–1164.

14. Gehan EA, Schneiderman MA. Historical and methodological developments in clinical trials at the National Cancer Institute. *Stat Med* 1990;9:871–880.
15. Halperin M, DeMets DL, Ware JH. Early methodological developments for clinical trials at the National Heart, Lung, and Blood Institute. *Stat Med* 1990;9:881–892.
16. Greenhouse SW. Some historical and methodological developments in early clinical trials at the National Institutes of Health. *Stat Med* 1990;9:893–901.
17. Byar DP. Discussion of papers on "historical and methodological developments in clinical trials at the National Institutes of Health." *Stat Med* 1990;9:903–906.
18. Organization, Review, and Administration of Cooperative Studies (Greenberg Report). A report from the Heart Special Project Committee to the National Advisory Heart Council, May 1967. *Control Clin Trials* 1988;9:137–148.
19. OPRR Reports. Code of Federal Regulations: (45 CFR 46) Protection of Human Subjects. National Institutes of Health, Department of Health and Human Services. Revised June 23, 2005. http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm.
20. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. *Fed Regist* 1979;44:23192–23197. http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.htm.
21. Nuremburg Code. http://www.hhs.gov/ohrp/references/nurcode.htm.
22. World Medical Association Declaration of Helsinki. http://www.wma.net/e/policy/b3.htm.
23. International Harmonised Tripartite Guideline: General Considerations for Clinical Trials: E8. December 17, 1997. http://www.fda.gov/downloads/RegulatoryInformation/Guidances/UCM129510.pdf.
24. Buoen C, Bjerrum OJ, Thomsen MS. How first-time-in-human studies are being performed: a survey of phase 1 dose-escalation trials in healthy volunteers published between 1995 and 2004. *J Clin Pharmacol* 2005;45:1123–1136.
25. Carbone PP, Krant MJ, Miller SP, et al. The feasibility of using randomization schemes early in the clinical trials of new chemotherapeutic agents:hydroxyurea (NSC-32065). *Clin Pharmacol Ther* 1965;6:17–24.
26. Anbar D. Stochastic approximation methods and their use in bioassay and phase I clinical trials. *Commun Stat Ser A* 1984;13:2451–2467.
27. Williams DA. Interval estimation of the median lethal dose. *Biometrics* 1986;42:641–645; correction in: *Biometrics* 1987;43:1035.
28. Storer B, DeMets D. Current phase I/II designs: are they adequate? *J Clin Res Drug Dev* 1987;1:121–130.
29. Storer B. Design and analysis of phase I clinical trials. *Biometrics* 1989;45:925–937.
30. Gordon NH, Willson JK. Using toxicity grades in the design and analysis of cancer phase I clinical trials. *Stat Med* 1992;11:2063–2075.
31. Schneiderman MA. Mouse to man: statistical problems in bringing a drug to clinical trial. *Proceedings of the 5th Berkeley Symposium of Math and Statistical Problems, University of California* 1967;4:855–866.
32. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics* 1990;46:33–48.
33. O'Quigley J, Chevret S. Methods for dose finding studies in cancer clinical trials: a review and results of a Monte Carlo Study. *Stat Med* 1991;10:1647–1664.
34. Wang O, Faries DE. A two-stage dose selection strategy in phase 1 trials with wide dose ranges. *J Biopharm Stat* 2000;10:319–333.
35. Babb J, Rogatko A. Bayesian methods for cancer phase I clinical trials. In: N. Geller (Ed.), *Advances in Clinical Trial Biostatistics*. New York: Marcel Dekker, 2004, pages 1–39.
36. Biswas S, Liu DD, Lee JJ, Berry DA. Bayesian clinical trials at the University of Texas M. D. Anderson Cancer Center. *Clin Trials* 2009;6:205–216.
37. Garrett-Mayer E. The continual reassessment method for dose-finding studies: a tutorial. *Clin Trials* 2006;3:57–71.

38. Babb J, Rogatko A, Zacks S. Cancer phase I clinical trials: efficient dose escalation with overdose control. *Stat Med* 1998;17:1103–1120.
39. Thall PF, Millikan RE, Mueller P, Lee S-J. Dose-finding with two agents in phase I oncology trials. *Biometrics* 2003;59:487–496.
40. Cheung YK, Chappell R. Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics* 2000;56:1177–1182.
41. Crowley J, Ankerst DP (Eds.), *Handbook of Statistics in Clinical Oncology* (Second ed.). Boca Raton, FL: Chapman and Hall/CRC, 2006.
42. Ting N (Ed.), *Dose Finding in Drug Development.* New York: Springer, 2006.
43. Gehan EA. The determination of the number of patients required in a follow-up trial of a new chemotherapeutic agent. *J Chron Dis* 1961;13:346–353.
44. Fleming TR. One-sample multiple testing procedures for phase II clinical trials. *Biometrics* 1982;38:143–151.
45. Herson J. Predictive probability early termination plans for phase II clinical trials. *Biometrics* 1979;35:775–783.
46. Geller NL. Design of phase I and II clinical trials in cancer: a statistician's view. *Cancer Invest* 1984;2:483–491.
47. Whitehead J. Sample sizes for phase II and phase III clinical trials: an integrated approach. *Stat Med* 1986;5:459–464.
48. Chang MN, Therneau TM, Wieand HS, Cha SS. Designs for group sequential phase II clinical trials. *Biometrics* 1987;43:865–874.
49. Simon R, Wittes RE, Ellenberg SS. Randomized phase II clinical trials. *Cancer Treat Rep* 1985;69:1375–1381.
50. Jung S, Carey M, Kim K. Graphical search for two-stage designs for phase II clinical trials. *Control Clin Trials* 2001;22:367–372.
51. Case LD, Morgan TM. Duration of accrual and follow-up for two-stage clinical trials. *Lifetime Data Anal* 2001;7:21–37.
52. Thall P, Simon R. Recent developments in the design of phase II clinical trials. In: P. Thall, (Ed.), *Recent Advances in Clinical Trial Design and Analysis*. Norwell, MA: Kluwer, 1995, pages 49–72.
53. Grieve AP, Krams M. ASTIN: a Bayesian adaptive dose-response trial in acute stroke. *Clin Trials* 2005;2:340–351.
54. Lee YJ, Staquet M, Simon R, et al. Two-stage plans for patient accrual in phase II cancer clinical trials. *Cancer Treat Rep* 1979;63:1721–1726.
55. Schaid DJ, Ingle JN, Wieand S, Ahmann DL. A design for phase II testing of anticancer agents within a phase III clinical trial. *Control Clin Trials* 1988;9:107–118.
56. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials* 1989;10:1–10.
57. Thall PF, Simon R. Incorporating historical control data in planning phase II clinical trials. *Stat Med* 1990;9:215–228.
58. Schmidli H, Bretz F, Racine-Poon A. Bayesian predictive power for interim adaptation in seamless phase II/III trials where the endpoint is survival up to some specified timepoint. *Stat Med* 2007;26:4925–4938.
59. Sylvester RJ, Staquet MJ. Design of phase II clinical trials in cancer using decision theory. *Cancer Treat Rep* 1980;64:519–524.
60. Berry D. Decision anlaysis and Bayesian methods in clinical trials. In: P. Thall (Ed.), *Recent Advances in Clinical Trial Design and Analysis.* Norwell, MA: Kluwer, 1995, pages 125–154.
61. Solomon SD, McMurray JJV, Pfeffer MA, et al. Cardiovascular risk associated with celecoxib in a clinical trial for colorectal adenoma prevention. *N Engl J Med* 2005;352:1071–1080.
62. Psaty BM, Furberg CD. COX-2 inhibitors – lessons in drug safety. *N Engl J Med* 2005;352:1133–1135.
63. Bolen S, Feldman L, Vassy J, et al. Systematic review: comparative effectiveness and safety of oral medications for type 2 diabetes mellitus. *Ann Intern Med* 2007;147:386–399.
64. The Digitalis Investigation Group. The effect of digoxin on mortality and morbidity in patients with heart failure. *N Engl J Med* 1997;336:525–533.

65. The Intermittent Positive Pressure Breathing Trial Group. Intermittent positive pressure breathing therapy of chronic obstructive pulmonary disease – a clinical trial. *Ann Intern Med* 1983;99:612–620.
66. Silverman WA. The lesson of retrolental fibroplasia. *Sci Am* 1977;236:100–107.
67. Echt DS, Liebson PR, Mitchell LB, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial. *N Engl J Med* 1991;324:781–788.
68. Alderson P, Roberts I. Corticosteroids for acute traumatic brain injury. *Cochrane Database Syst Rev* 2000;(2):CD000196.
69. Roberts I, Yates D, Sandercock P, et al. Effect of intravenous corticosteroids on death within 14 days in 10008 adults with clinically significant head injury (MRC CRASH trial): randomised placebo-controlled trial. *Lancet* 2004;364:1321–1328.
70. Edwards P, Arango M, Balica L, et al. Final results of MRC CRASH, a randomised placebo-controlled trial of intravenous corticosteroid in adults with head injury – outcomes at 6 months. *Lancet* 2005;365:1957–1959.
71. Alderson P, Roberts I. Corticosteroids for acute traumatic brain injury. *Cochrane Database Syst Rev* 2005;(1):CD000196.
72. Stone GW, Lansky AJ, Pocock SJ, et al. Paclitaxel-eluting stents versus bare-metal stents in acute myocardial infarction. *N Engl J Med* 2009;360:1946–1959.
73. James SK, Stenestrand U, Lindbäck J, et al. Long-term safety and efficacy of drug-eluting versus bare-metal stents in Sweden. *N Engl J Med* 2009;360:1933–1945.
74. Byar DP, Schoenfeld DA, Green SB, et al. Design considerations for AIDS trials. *N Engl J Med* 1990;323:1343–1348.
75. Levine C, Dubler NN, Levine RJ. Building a new consensus: ethical principles and policies for clinical research on HIV/AIDS. *IRB* 1991;13:1–17.
76. Spiers HR. Community consultation and AIDS clinical trials, part I. *IRB* 1991;13:7–10.
77. Emanuel EJ, Grady C. Four paradigms of clinical research and research oversight. In: E.J. Emamuel, C. Grady, R.A. Crouch, R.K. Lie, F.G. Miller, D. Wendler (Eds.), *The Oxford Textbook of Clinical Research Ethics.* Oxford: Oxford University Press, 2008, pages 222–230.
78. Abigail Alliance for Better Access to Developmental Drugs. http:www.abigail-alliance.org.
79. Furberg CD. The impact of clinical trials on clinical practice. *Arzneim-Forsch./Drug Res* 1989;39:986–988.
80. Lamas GA, Pfeffer MA, Hamm P, et al. Do the results of randomized clinical trials of cardiovascular drugs influence medical practice? *N Engl J Med* 1992;327:241–247.
81. Friedman L, Wenger NK, Knatterud GL. Impact of the Coronary Drug Project findings on clinical practice. *Control Clin Trials* 1983;4:513–522.
82. Boissel JP. Impact of randomized clinical trials on medical practices. *Control Clin Trials* 1989;10:120S–134S.
83. Rosenberg Y, Schron E, Parker A. How clinical trial results are disseminated: use and influence of different sources of information in a survey of US physicians. *Control Clin Trials* 1994;15:16S.
84. Schron E, Rosenberg Y, Parker A, Stylianou M. Awareness of clinical trials results and influence on prescription behavior: a survey of US physicians. *Control Clin Trials* 1994;15:108S.
85. Ayanian JZ, Haustman PJ, Guadagnoli E, et al. Knowledge and practices of generalist and specialist physicians regarding drug therapy for acute myocardial infarction. *N Engl J Med* 1994;331:1136–1142.
86. Ford ES, Ajani UA, Croft JB, et al. Explaining the decrease in U.S. deaths from coronary disease, 1980–2000. *N Engl J Med* 2007;356:2388–2398.
87. Casper M, Wing S, Strogatz D, Davis CE, Tyroler HA. Antihypertensive treatment and US trends in stroke mortality, 1962–1980. *Am J Public Health* 1992;82:1600–1606.
88. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;342:1878–1886.
89. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;342:1887–1892.

90. Bjelakovic G, Nikolova D, Gluud LL, Simonetti RG, Gluud C. Mortality in randomized trials of antioxidant supplements for primary and secondary prevention: systematic review and meta-analysis. *JAMA* 2007;297:842–857.
91. Hulley S, Grady D, Bush T, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *JAMA* 1998;280:605–613.
92. Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women. *JAMA* 2002;288:321–333.
93. The Women's Health Initiative Steering Committee. Effects of conjugated equine estrogen in postmenopausal women with hysterectomy. *JAMA* 2004;291:1701–1712.
94. Furberg BD, Furberg CD. *Evaluating Clinical Research: All that Glitters is not Gold*. (Second ed.). New York: Springer, 2007.
95. Chalmers TC. Randomization of the first patient. *Med Clin North Am* 1975;59:1035–1038.
96. Spodick DH. Randomize the first patient: scientific, ethical, and behavioral bases. *Am J Cardiol* 1983;51:916–917.
97. Bonchek LI. Sounding board: are randomized trials appropriate for evaluating new operations? *N Engl J Med* 1979;301:44–45.
98. Van der Linden W. Pitfalls in randomized surgical trials. *Surgery* 1980;87:258–262.
99. Rudicel S, Esdail J. The randomized clinical trial in orthopaedics: obligation or option? *J Bone Joint Surg* 1985;67:1284–1293.
100. Murphy ML, Hultgren HN, Detre K, et al. Treatment of chronic stable angina – a preliminary report of survival data of the randomized Veterans Administration cooperative study. *N Engl J Med* 1977;297:621–627.
101. Takaro T, Hultgren HN, Lipton MJ, Detre KM. The VA cooperative randomized study of surgery for coronary arterial occlusive disease. 11. Subgroup with significant left main lesions. *Circulation* 1976;54:111–107.
102. Detre K, Peduzzi P, Murphy M, et al. Effect of bypass surgery on survival in patients in low- and high-risk subgroups delineated by the use of simple clinical variables. *Circulation* 1981;63:1329–1338.
103. Proudfit WL. Criticisms of the VA randomized study of coronary bypass surgery. *Clin Res* 1978;26:236–240.
104. Chalmers TC, Smith H Jr, Ambroz A, et al. In defense of the VA randomized control trial of coronary artery surgery. *Clin Res* 1978;26:230–235.
105. CASS Principal Investigators and Their Associates. Myocardial infarction and mortality in the Coronary Artery Surgery Study (CASS) randomized trial. *N Engl J Med* 1984;310:750–758.
106. Strachan CJL, Oates GD. Surgical trials. In: F.N. Johnson, S. Johnson (Eds.), *Clinical Trials*. Oxford: Blackwell Scientific, 1977.
107. Bunker JP, Hinkley D, McDermott WV. Surgical innovation and its evaluation. *Science* 1978;200:937–941.
108. Weil RJ. The future of surgical research. *PLoS Med* 2004;1:e13. doi:10.1371/journal.pmed.0010013.
109. Cook JA. The challenges faced in the design, conduct and analysis of surgical randomised controlled trials. *Trials* 2009. 10:9. doi:10.1186/1745-6215-10-9.
110. Chalmers TC, Sacks H. Letter to the editor: randomized clinical trials in surgery. *N Engl J Med* 1979;301:1182.
111. Greene HL, Roden DM, Katz RJ, et al. The cardiac arrhythmia suppression trial: first CAST…then CAST-II. *J Am Coll Cardiol* 1992;19:894–898.
112. ClinicalTrials.gov. http://clinicaltrials.gov/.
113. WHO International Clinical Trials Registry Platform. http://www.who.int/ictrp/network/en/.
114. Clinical Trials Registration in ClinicalTrials.gov (Public Law 110-85): Competing Applications and Non-Competing Progress Reports. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-023.html.
115. Notice of public process for the expansion of the Clinical Trials.gov registry and availability of a basic results database. *Fed Regist* 73(99). May 21, 2008. http://edocket.access.gpo.gov/2008/E8-11042.htm.

Chapter 2
Ethical Issues

People have debated the ethics of clinical trials for as long as trials have been conducted. The arguments have changed over the years and perhaps become more sophisticated, but many of them involve issues such as the physician's obligations to the individual patient versus societal good, clinical equipoise, study design considerations such as randomization and the choice of control group, including use of placebo, informed consent, conduct of trials in underdeveloped areas, conflict of interest, participant confidentiality and sharing of data and specimens, and publication bias.

A well-designed trial should answer important public health questions without impairing the welfare of individuals. There may, at times, be conflicts between a physician's perception of what is good for his or her patient and the design and conduct of the trial. In such instances, the needs of the patient must predominate.

Ethical issues apply in all stages of a clinical trial. In this chapter, we summarize some of the major factors involving ethics in design, conduct, and reporting of clinical trials. As will be noted, several of the issues are unsettled and have no easy solution. We expect, however, that investigators will at least consider these issues in the planning stages of trials so that high ethical standards can be applied to all trials.

Emanuel et al. [1] listed seven criteria that they considered essential to the ethical conduct of clinical research. These criteria are value, scientific validity, fair selection of participants, favorable benefit/risk balance, independent review, informed consent, and respect for participants. Independent review is generally conducted by ethics review committees specifically constituted for oversight of research with human subjects. In the United States, such committees are termed Institutional Review Boards (IRBs). Other names used outside the US are Research Ethics Committees, Ethics Committees, or Ethics Review Committees. Although the role of ethics review committees is discussed later in this chapter under Informed Consent, it must be emphasized that independent review by these committees and others, such as data and safety monitoring boards, applies to all aspects of a trial.

We encourage the reader to seek out any of the many books and journals devoted to ethical aspects of clinical research. Those go into the issues, including ones we do not address, in considerable depth. A particularly relevant book is the Oxford Textbook of Clinical Research Ethics, many chapters of which relate directly to clinical trials [2]. The reader is also referred to several key documents [3–6].

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_2, © Springer Science+Business Media, LLC 2010

Fundamental Point

Investigators and sponsors of clinical trials have ethical obligations to trial participants and to science and medicine.

Planning and Design

Does the Question Require a Clinical Trial?

An early decision relates to whether a clinical trial is even necessary. Not all questions need to be answered, and not all of those that should be answered require clinical trials. Sometimes, other kinds of clinical studies may be able to address the question at least as well as, or even better than, a clinical trial. Even if the answer may not be quite as good, the added benefits from the trial may not be worth the added risk.

Because clinical trials involve administering something (drug, device, biologic, or procedure) to someone, or attempting to change someone's behavior, there may be adverse as well as the hoped-for positive results. Although some of the potential adverse consequences may be known before the trial is started, and therefore prevented or minimized, others may arise unexpectedly during the trial or be more serious than anticipated. The question being addressed by the clinical trial, therefore, must be important enough to justify the possible adverse events. The question must have relevant clinical, public health, and/or other scientific value. A trivial question should not expose study participants to risk of harm, either physical or emotional. Harm can be either a direct result of the intervention or indirect, from withholding something beneficial. The study investigator, sponsor or funder, and institutions where the study will be performed must all ensure that the question is sufficiently important and the trial is appropriately conducted to justify those risks. Otherwise, the adage "Above all, do no harm," applies.

Though the question may be important, the clinical trial may be infeasible or unethical. An obvious example is cigarette smoking. Performing clinical trials in nonsmokers, providing half of them with cigarettes, to prove that smoking is harmful is clearly unethical. Observational studies have given us sufficient evidence to answer that question. The Cardiac Arrhythmia Suppression Trial (CAST) [7] was designed to determine if suppression of ventricular arrhythmias with antiarrhythmic agents in people with heart disease would lead to a reduction in sudden cardiac death. After two of the three antiarrhythmic drugs were seen to be harmful and stopped, some asked whether the study might be continued, but reconfigured to demonstrate that quinidine, a long-used drug with some properties similar to the two discontinued agents, would also be harmful. The CAST investigators quickly decided that designing a trial specifically to prove harm, especially serious harm, would be unethical. Although the outcome of a trial is uncertain, the primary response variable should always be one where either benefit or noninferiority is potentially achievable.

Two kinds of trials raise ethical issues because of concerns about the balance between potential benefits to society (and perhaps to participants) and the risks of harm and discomfort to participants. In both, the likelihood of immediate benefit to the study participants exists, but is remote. One involves “me too” or “marketing” (also termed “seeding”) trials. Such clinical trials are conducted to show that a new drug or new version of an old drug is at least as good as (noninferior to) a drug already proven to be beneficial. Other than enhancing the financial status of the industry sponsor, there may be little benefit to the new drug. Yet trial participants are being put at risk from a drug with unknown adverse effects, some of which might be serious. If the new drug has some potential improvement over the existing one, the trial might be justified. Perhaps the new drug is easier to take (e.g., once a day rather than twice a day administration), is better tolerated, or causes fewer adverse events. One could also argue that having more than one drug with similar benefits is good for the economy, fostering lower medical care costs. But in the end, those conducting such trials should show how the question is important and how there will be meaningful benefits for patients.

A second kind of trial, the ethics of which have been debated, is the early phase study. If these studies are performed in healthy volunteers, there is a nontrivial chance that they will be harmed, but have no opportunity to benefit, other than from whatever payment they receive as a result of their participation. Some people regularly enroll in such studies for the payment [8]. It has been argued that with proper attention to study design and safety monitoring, appropriate evaluation by ethics review committees, and true informed consent, these studies are ethical [9]. As always, risk must be kept to a minimum and the payment must not be so great as to encourage participants to do something that would place them at serious risk. The pros and cons of various payment models for research participants are discussed by Dickert and Grady [10]. As with other clinical research, early phase studies are only ethical if investigators and sponsors do whatever is necessary to minimize risk. Unfortunately, instances when investigators may not have taken proper care have occurred and received widespread attention [11–13].

Some early phase studies are conducted with participants who have a disease or condition. Patients with cancer that has not responded to other therapies may volunteer for such trials, hoping that the experimental intervention will prove beneficial. Given the small size of these studies and the unfortunate fact that most interventions early in their development do not prove beneficial, some have even questioned the ethics of these trials. But even if there is only a slight possibility of improvement, as long as there is adequate informed consent and the expectation of benefit to society from the knowledge to be gained, most would agree that these trials can be conducted in an ethical manner [14, 15].

Randomization

In the typical “superiority trial” described in Chap. 5, randomization is usually done on top of standard or usual therapy, which all participants should receive. (The special

issues related to noninferiority trials are discussed in Chap. 5.) Randomization has often been a problem for physicians and other clinicians who believe they must be able to convey to their patients a treatment course of action. The researcher, however, must accept uncertainty. Therefore, an objection to random assignment should only apply if the investigator believes that a superior therapy exists. If that is the case, she should not participate in a trial that involves the preferred therapy. On the other hand, if she truly cannot say that one treatment is better than another, there should be no ethical problem with randomization. Such judgments regarding efficacy obviously vary among investigators. Because it is unreasonable to expect that an individual investigator has no preference, not only at the start of a trial but also during its conduct, the concept of "clinical equipoise" was proposed [16]. In this concept, the presence of uncertainty as to the benefits or harm from an intervention among the expert medical community, rather than in the individual investigator, is a justification for a clinical trial. Some have maintained that until an intervention has been proven beneficial, randomization is the most ethical approach and one that will provide the correct answer soonest [17–20].

Control Group

Choice of the control group is a major design issue in all clinical trials. If there is a known best therapy, one would generally expect the new intervention to be compared with that therapy, or added to it. But the optimal therapy may not be widely used for various reasons. These could include cost, unavailability of the therapy or lack of sufficient clinicians competent to administer it, lack of acceptance by the practicing clinical community, socioeconomic and cultural differences, or other factors. Depending on these circumstances, some trials may not use the best known therapy or standard of care as the control. They may rely on what the practicing communities typically do, or usual therapy [21]. Investigators and ethics review committees need to judge whether the usual therapy deprives participants of a proven better treatment that they would otherwise receive. If so, serious ethical concerns arise. A major area of disagreement has been the degree of responsibility of investigators to ensure that all participants receive the best proven therapy as a control or background care, even if usual care in the community in which the trial is being conducted is not up to that standard [22]. (See also the section below, Trials in Developing Countries.)

Considerable confusion has arisen when people talk about placebo-controlled trials, as they may refer to different kinds of designs. Often, a new intervention is added to usual care or standard care, and compared against that care plus placebo. Sometimes, a new intervention is seen as a possible replacement for an existing therapy, yet for various reasons, it is not thought appropriate to compare the new intervention against the existing therapy. The commonly used therapy, for example, may not have been proven to be beneficial, or it may be poorly tolerated. Therefore, a placebo comparator is used instead of the existing therapy.

Even if a proven therapy exists, whether short-term discontinuation of that therapy for the purpose of conducting a placebo-controlled trial is harmful depends

on the condition being studied. Exposing participants to serious harm by withholding beneficial treatment is unethical even in the short term. For conditions causing only mild to moderate discomfort, it may be acceptable. For example, investigators evaluating new analgesic agents might choose to use a placebo control, as long as any pain or discomfort is treated promptly. As always, there will be borderline cases that require discussion and review by ethics review committees [23].

Freedman et al. [24, 25] acknowledged that many factors enter into a decision regarding the use of a placebo control. They argued that if an accepted treatment exists, much of the time a placebo control is unethical and, indeed, unnecessary. Rothman and Michels [26, 27] also maintained that in many cases, a placebo has been used inappropriately because a proven therapy existed. This debate occurred with the Enhanced Suppression of the Platelet IIb/IIIa Receptor with Integrilin Trial (ESPRIT) [28–30]. The decision to use a placebo control, rather than another proven IIb/IIIa receptor inhibitor, was only allowed after it was shown that many cardiologists were not persuaded by the prior evidence. We think that this is a valid argument only if all investigators (including referring clinicians) have been informed about the current evidence and make the decision to conduct another placebo-controlled trial because they question the applicability of that evidence. Ethics review committees must have full knowledge, and informed consent must contain the relevant information.

Whenever an investigator considers using a placebo control, she must assess whether it will provide the most meaningful answer to the question being addressed, and will not cause serious harm. Importantly, all participants must be told that there is a specified probability, e.g., 50%, of their receiving placebo. The World Medical Association Declaration of Helsinki [5], the Council for International Organizations of Medical Sciences (CIOMS) [6], regulatory bodies [31], and others have guidelines for the use of placebo. Miller summarizes the issues that must be considered by investigators [32].

Protection from Conflict of Interest

A widely expressed concern in much clinical research is the potential for conflict of interest on the part of the investigators. In the context of ethical issues, conflict of interest can lead to bias in design, conduct, analysis, interpretation, and communication of findings. Conflict of interest is generally considered in the financial context, but intellectual or other conflicts may also occur [33]. Ideally, no investigator would have any interests other than the well-being of the study participants and the generation of new knowledge that will improve clinical care and public health. That is unrealistic, however, given that most investigators receive research funding from government, industry, or others with considerable interest in the outcome of the study. Many investigators have also spent a career attempting to advance the science and could be disappointed if their favorite theory turns out to be incorrect. Therefore, most clinical trials find it easier to manage conflict of interest than to avoid it completely.

The role of disclosure of financial relationships to participants and others has been reviewed and recommendations proposed [34]. Among these recommendations, it was noted that because many participants may not fully appreciate the impact that financial relationships might have on research design, conduct, and analysis, in addition to requiring disclosure, IRBs and others should "play a significant role in determining the acceptability of these relationships" [34]. We think that disclosure and IRB or other oversight may be sufficient for early phase studies. It may not be sufficient, however, for late phase trials: those that are designed to have major implications for clinical practice. Most clinical trials are sponsored by industry, and although the investigators enrolling and following participants may not stand to gain financially from the results of the trial, the sponsors clearly do. Therefore, all data collection and analysis should be conducted by groups independent of the industry sponsor. Ideally, this should also occur in trials sponsored by others. Any investigators who have economic interests in the outcome either should not participate or should not have opportunities to affect and publish the trial outcome. This may mean that the lead investigator in multi-investigator studies or the investigator in single investigator studies should have no conflicts if the study is one likely to change practice. Other key investigators with major conflicts should also be barred from such trials. If the investigators have limited roles or only small financial investments, it may be acceptable for them to participate. We recognize that the situation is more complicated when those designing and overseeing, and perhaps coauthoring publications, are employees of the company sponsoring the trial. Nevertheless, complete openness and data analysis by an independent group remain important. The use of external independent oversight bodies and clear lines of authority may mitigate conflict of interest. In the end, however, clinical trial results must be believed and accepted by the clinical communities. To the extent that conflict of interest (real or perceived) lessens that acceptance, the study is impaired. Therefore, all appropriate ways of minimizing and managing conflicts should be used.

Informed Consent

Proper informed consent is essential. Partly as a result of terrible things done in the name of clinical research, various bodies developed guidelines such as the Nuremberg Code [4], the Declaration of Helsinki [5], the Belmont Report [3], and the International Ethical Guidelines for Biomedical Research Involving Human Subjects [6]. These guidelines lay out standards for informed consent that are commonly followed internationally. In the USA, in parallel to the Belmont Report, the United States Congress passed laws that require adherence to informed consent regulations by those receiving government support – the so-called Common Rule, or 45 CFR 46 [35] – and those evaluating agents under the auspices of the Food and Drug Administration [36]. These regulations require that clinical research studies be reviewed by IRBs and establish the membership and procedures that IRBs must follow.

One of the primary roles of the IRB is to ensure that there is true, voluntary informed consent. The Common Rule requires consent forms to contain basic elements. Table 2.1 lists these, as well as other elements that may be added as appropriate. Simply adhering to legal requirements does not ensure informed consent [37–39]. Informed consent is a process that can take considerable time and effort; it is not simply a matter of getting a form signed. In many, perhaps most, clinical trial settings, true informed consent can be obtained. Potential participants have the capacity to understand what is being requested of them, they have adequate time to consider the implications of joining a trial, to ask questions, and to take information

Table 2.1 Informed consent checklist – basic and additional elements [35]

A statement that the study involves research
An explanation of the purposes of the research
The expected duration of the subject's participation
A description of the procedures to be followed
Identification of any procedures which are experimental
A description of any reasonably foreseeable risks or discomforts to the subject
A description of any benefits to the subject or to others which may reasonably be expected from the research
A disclosure of appropriate alternative procedures or courses of treatment, if any, that might be advantageous to the subject
A statement describing the extent, if any, to which confidentiality of records identifying the subject will be maintained
For research involving more than minimal risk, an explanation as to whether any compensation, and an explanation as to whether any medical treatments are available, if injury occurs and, if so, what they consist of, or where further information may be obtained
An explanation of whom to contact for answers to pertinent questions about the research and research subjects' rights, and whom to contact in the event of a research-related injury to the subject
A statement that participation is voluntary, refusal to participate will involve no penalty or loss of benefits to which the subject is otherwise entitled, and the subject may discontinue participation at any time without penalty or loss of benefits, to which the subject is otherwise entitled
Additional elements, as appropriate
A statement that the particular treatment or procedure may involve risks to the subject (or to the embryo or fetus, if the subject is or may become pregnant), which are currently unforeseeable
Anticipated circumstances under which the subject's participation may be terminated by the investigator without regard to the subject's consent
Any additional costs to the subject that may result from participation in the research
The consequences of a subject's decision to withdraw from the research and procedures for orderly termination of participation by the subject
A statement that significant new findings developed during the course of the research, which may relate to the subject's willingness to continue participation, will be provided to the subject
The approximate number of subjects involved in the study

home to review and discuss with their families and personal physicians, and they are familiar with the concepts of research and voluntary consent. As discussed in the Privacy and Confidentiality section below, investigators may share data and biospecimens with other researchers. If such sharing is planned or required by the sponsor, the informed consent must make it clear that sharing will occur and that the data may be used for purposes other that those of the trial for which the person is volunteering.

Sometimes, people who are ill may not understand that a clinical trial is a research endeavor. They may believe that they are receiving therapy for their condition. This may happen in early phase trials of new drugs that are being developed for serious, untreatable diseases, or in any clinical trial testing, a promising intervention for a serious or chronic condition. Patients may view the trial as the last or best possibility for cure. Sometimes, clinicians are also researchers, and may seek to enroll their own patients into clinical trials. These situations can lead to what has been termed "therapeutic misconception" [40]. The distinction between research, in essence an experiment, and clinical care may blur. We do not advocate preventing clinicians from enrolling their own patients into clinical trials. However, extra effort must be made to provide the patients with the information needed to judge the merits of volunteering to enter the research, separate from their clinical care.

The situations where participant enrollment must be done immediately, in comatose patients, or in highly stressful circumstances and where the prospective participants are minors or not fully competent to understand the study are more complicated and may not have optimal solutions. In the U.S., FDA [41] and the Department of Health and Human Services [42] guidelines allow for research in emergency situations, when informed consent is not possible. Under these regulations, IRBs may approve the study as long as a series of special conditions has been met, including that there has been community consultation and a safety committee is formed to monitor accumulating data. Similar research is also allowed in Canada [43] and under the European Medicines Agency (EMA) Guidelines for Good Clinical Practice [44]. A trial of thrombolytics versus placebo in the context of resuscitation for cardiac arrest was successfully conducted under the EMA guidelines [45]. In this trial, local ethics committees agreed that the trial could be done without informed consent prior to enrollment. Instead, consent was later given by surviving participants or their family members or others.

Some have questioned all research in emergency settings because of the lack of prior informed consent, and several such clinical trials have been quite controversial. An example is a trial of a product intended to be used as a blood substitute in trauma patients [46]. Because patients were unconscious at the time of administration of the blood substitute, consent could not be obtained. Therefore, community consultation was obtained before local IRBs approved the study. However, there were allegations that safety problems noted in earlier trials of the agent were not published or otherwise disclosed to those bodies. We do not take a position on the merits of this particular trial, and we support the concept of being able to conduct important research in settings where full informed consent before enrollment is not possible. The sponsors and investigators, though, must be completely open about all data relevant to the conduct of such studies and must follow all local regulations [47].

Failure to do so harms not only the unwitting participants but also the entire field of research in emergency settings.

Also contentious is the role of consent from participant surrogates when the study participant is unable to provide fully informed consent. This typically happens with research in minors, when parents or other guardians make the decisions. Special review is required for pediatric research; requirements vary depending on the expected risks from the study [35]. Other situations, such as research in emotionally or mentally impaired individuals also have generated discussion and guidelines regarding use of surrogate consent [48, 49]. Less clear is the use of surrogate consent for potential study participants who are temporality unable to understand the nature of the study and give consent. This issue arose in research in people with the acute respiratory distress syndrome [50]. Suggestions for accommodating research in such situations include risk assessment, determination of patient capacity, and reconsent [51]. As in all such situations, judgment on the part of investigators, sponsors, IRBs, and others will be required and second-guessing will inevitably occur.

Conduct

Trials in Developing Countries

Many clinical trials are international. The ability to enroll and follow participants in more than one country assists in enrollment and may assist in generalizing the results of the trial to different populations and settings. However, trials that are conducted in developing areas raise ethical issues. Are they conducted in those regions because the disease of interest is prevalent there, and the results relevant to the region? Or are the countries or regions selected primarily for convenience, low cost, or fewer administrative and regulatory burdens? The control group may be receiving less than optimal care, and thus may have a higher event rate, permitting a smaller, shorter, and less expensive trial. If the trial is conducted for those reasons, it is unethical. Some have said that the investigators are obligated to ensure that all participants receive care that is optimal without regard to usual practice in the country where the trial is being conducted. Others have maintained that it is sufficient if the participants receive care at least as good as what they would receive had they not been in the trial. This was the argument of the investigators in the Vietnam Tamoxifen Trial of adjuvant oophorectomy and tamoxifen in treatment of breast cancer. State of the art treatment by US standards (including radiation) was not available and not likely to be available. What was being tested was whether a simple and affordable treatment like tamoxifen would be better than what was available [52].

Extrapolation of study results from less developed regions to highly developed countries with very different health care systems and standards of care, and vice versa, has also been questioned. Some studies suggest that the outcomes may indeed be different [53, 54].

After the trial ends, what is the obligation of the investigators to provide an intervention shown to be beneficial, both to the study participants and to the broader population? This and other issues have no easy answers. We believe, however, that trials should only be conducted in places and with participants likely to benefit from the results and with informed consents that clearly describe what will be done at the end of the trial. The results from the trial must be able to be applied to clinical practice in the population from which the participants came [55].

Recruitment

Recruitment of trial participants is often one of the more challenging aspects of conducting a clinical trial (see Chap. 10). Unless an adequate number of participants is entered, the trial will not be able to answer the questions about benefit and risk. Therefore, there is great pressure to recruit an adequate number of participants and to do so as quickly as possible. The use of some financial incentives, such as "finder's fees," i.e., payment to physicians for referring participants to a clinical trial investigator, is inappropriate, in that it might lead to undue pressure on a prospective participant [56]. This differs from the common and accepted practice of paying investigators a certain amount for the cost and effort of recruiting each enrolled participant. Even this practice becomes questionable if the amount of the payment is so great as to induce the investigator to enroll inappropriate participants [10].

Study participants may be paid for their involvement in clinical trials. Typically, payment is meant to compensate them for the time, effort, and expense of attending clinic visits. Studies that enroll healthy volunteers (usually phase I trials) will often provide payment beyond reimbursement for expenses. The amount generally depends on the time required and the amount of pain and risk involved in any procedures. As with paying investigators, when the amount is such that people, whether they are healthy volunteers or patients, might make unwise or dangerous decisions, it becomes excessive. Participants should never be paid more for taking on more risk. Ethics review committees often have guidelines as to appropriate amounts for various kinds of studies and procedures and must ensure that the amount provided does not create an undue influence.

As discussed in Chap. 9, many potentially eligible trial participants may be on medication. This treatment may be for the condition that will be studied or some other reason. In order to assess the participant's condition at baseline, the investigator may be tempted to withdraw medication, at least temporarily. For example, one might be interested in enrolling people at high risk of cardiovascular disease, and thus try to accrue those with hypertension. But an accurate baseline blood pressure might not be obtainable in those already on treatment. It might not even be clear that the participant already on antihypertensive drugs would have met the eligibility criteria if not on medication. Should one withdraw the drug or simply accept that those on treatment probably truly had hypertension, especially if on treatment they still have high normal blood pressures? Usually, the latter is the better course of action.

Safety and Efficacy Monitoring

Occasionally, during a trial, important information relevant to informed consent derives either from other studies or from the trial being conducted. In such cases, the investigator is obligated to update the consent form and notify current participants in an appropriate manner. A trial of antioxidants in Finnish male smokers (the Alpha-Tocopherol Beta Carotene Prevention Study) indicated that beta carotene and vitamin E may have been harmful with respect to cancer or cardiovascular diseases, rather than beneficial [57]. Because of those findings, investigators of the ongoing Carotene and Retinol Efficacy Trial (CARET) informed the participants of the results and the possible risks [58]. CARET was subsequently stopped earlier than planned because of adverse events similar to those seen in the Finnish trial. The investigator of a third trial of antioxidants, the Age-Related Eye Disease Study (AREDS) then notified their participants of the findings from both the Finnish study and CARET [59, 60].

Five trials of warfarin in patients with atrial fibrillation were being conducted at approximately the same time [61]. After the first three ended, showing clear benefit from warfarin in the reduction of strokes, the remaining two found it difficult ethically to continue. Interim results from the Heart and Estrogen/Progestin Replacement Study (HERS) [62] and the Women's Health Initiative (WHI) [63] evaluation of estrogen suggested that thromboembolic adverse events that had not been clearly presented in the informed consent were occurring. In both studies, the data and safety monitoring boards debated whether the studies should stop or continue with additional actions taken. The trials continued, but participants in those trials and medical communities were notified of these interim findings [64, 65]. Not only is such a practice an ethical stance, but a well-informed participant is usually a better trial participant. How much data should be provided to study participants and when, and the role of independent safety monitoring groups in this decision, are still areas of debate [66].

The issue of how to handle accumulating data from an ongoing trial is a difficult one, and is further discussed in Chap. 16. With advance understanding by both participants and investigators that they will not be told interim results unless they show clear benefit or harm, and that there is a responsible safety monitoring group, ethical concerns should be lessened, if not totally alleviated.

Early Termination for Other than Scientific or Safety Reasons

Clinical trials are only ethical if there are adequate resources to conduct them and see them to completion. Trials may (and should) be stopped early if there are safety concerns or if there are scientific reasons to do so (see Chap. 15). It is inappropriate, however, to stop a trial early because the sponsor changes its mind about marketing priorities or failed to adequately plan for sufficient resources. In such cases, participants who had enrolled did so with the understanding that they would be

helping to advance medical knowledge. In the process, they put themselves at possibly considerable risk. To fail to complete the study is a serious breach of ethics. An example when this happened is the Controlled Onset Verapamil Investigation of Cardiovascular End Points (CONVINCE) trial [67]. Partway through follow-up, the sponsor ended the study for other than scientific or safety reasons. As noted in an editorial by Psaty and Rennie [68], "the responsible conduct of medical research involves a social duty and a moral responsibility that transcends quarterly business plans…"

In another situation, an investigator with inadequate funds to complete his trial solicited money from participants in the trial so that he could continue purchasing the experimental drug [69]. Because the trial was being conducted in patients with a fatal condition, amyotrophic lateral sclerosis, the study participants viewed the trial as a last hope and were therefore under considerable pressure to donate. We view such actions as completely unethical. Plans for conducting the trial, including obtaining experimental agents, must be in place before the trial begins.

With all trials, investigators need to plan in advance how they will handle end of study issues such as whether participants will have continued access to the intervention and transition to appropriate medical care.

Privacy and Confidentiality

The issues of privacy and confidentiality have received considerable attention. The widespread uses of electronic media have made many people concerned about the privacy of their medical records, including research records. Electronic medical records have simplified the tasks of finding potentially eligible participants for trials, conducting international multicenter studies, following up on participants during and after the studies, and sharing data with other researchers. They have also led to laws restricting what kinds of medical records can be shared and with whom, in the absence of clear permission from the patients. In the U.S., the Health Insurance Portability and Accountability Act (HIPAA) primarily addresses privacy issues in clinical practice. However, there are clinical research provisions that affect how investigators identify, contact, and obtain informed consent from prospective participants, and how study data are maintained and provided to others [70] (see also Chap. 10). These laws, in turn, have generated articles pointing out the increased difficulty in conducting clinical research. Policies encouraging or mandating sharing of data and biospecimens from research studies [71–73] may conflict with the objectives of maintaining confidentiality. If data are shared with other researchers for unspecified purposes, might participants who volunteered for a trial object to their data being used for goals of which they might not approve? If the original informed consent does not allow for use of the biospecimens by others or for purposes different from the stated ones, either the biospecimens cannot be shared or new informed consents must be obtained. The increasing availability and use of genetic material adds to this conflict. Fear of employment or health insurance discrimination based on genetic information may make some people unwilling to participate in trials

if complete confidentiality cannot be ensured. It is probably not possible to share data and specimens that are useful to the recipient investigator while also maintaining perfect deidentifiability. Some compromises are inevitable. At the current time, there are no clear solutions to these issues, but trial participants must have a right to make informed choices. Clinical trial investigators need to be aware of the concerns, and to the extent possible, plan to address them before the study starts.

Data Falsification

There has been concern about falsification of data and entry of ineligible, or even phantom participants in clinical trials (see Chap. 10). A case of possible falsification that gained considerable attention was a trial of bone morphogenetic protein-2 in the management of fractures due to combat injuries [74]. An editorial in the journal that published the article, which had purported to show benefit from treatment, said that "much of the paper was essentially false" and announced the article's withdrawal [75]. A trial of lumpectomy and radiation therapy for breast cancer was severely harmed because of falsified data on a small number of participants at one of many enrolling sites. The overall results were unchanged when the participants with the falsified data were not included [76, 77]. Nevertheless, the harm done to the study and to clinical trials in general was considerable. We condemn all data fabrication. It is important to emphasize that confidence in the integrity of the trial and its results is essential to every trial. If, through intentional or inadvertent actions, that confidence is impaired, not only have the participants and potentially others in the community been harmed, the trial loses its rationale, which is to influence science and medical practice. Chapter 11 reviews issues of ensuring data quality.

Reporting

Publication Bias, Suppression, and Delays

All investigators have the obligation to report trial results fully and in a timely fashion. As discussed in Chap. 19, it is well known that publication bias exists. Positive or exciting findings are more likely to be published than null results. In one survey of 74 trials of antidepressant agents, 38 were considered to have results favorable to the intervention. All but one of these were published. Of the 36 studies considered not to have favorable results, 22 were not published. Eleven others were published in ways that obscured the lack of favorable results [78]. Heres and colleagues examined trials of head-to-head comparisons of second generation antipsychotic agents [79]. Ninety percent of the trials sponsored by industry were reported in favor of the sponsor's drug. Interestingly, this occurred even with trials that compared the same drugs – but the outcome changed when the sponsor was a different company.

It is more probable that large, late phase trials will be published regardless of the results than will small, early stage trials. There are exceptions, however. As discussed in Chap. 5, the results of the second Prospective Randomized Amlodipine Survival Evaluation trial (PRAISE-2), although presented, were not published. The same is undoubtedly true of other trials with disappointing outcomes.

An important advance in ensuring publication is the requirement by many journals [80], sponsors such as the US National Institutes of Health [81], and the US Food and Drug Administration [82] that trials be registered at initiation in one of the accepted registration sites. Although it is not a complete solution to the problem of failure to make public the results of all trials, registration allows for easier tracking of trials that are initiated, but perhaps never completed or never published.

We take the position that the results of all clinical trials should be published regardless of the findings. It is important that the totality of the information, pro and con, be available so that those designing other studies and clinicians can make informed decisions. If the study results are not published, it is also unfair to the participants who volunteered for a trial with the understanding that they would be helping medical research. The so-called "gag clauses" in industry-sponsored trials [83] are both antithetical to academic freedom and contrary to ethical practice.

Conflicts of Interest and Publication

All researchers have biases of some sort. It is understandable that an investigator's perspective will enter into a publication, even though best efforts are made to be objective in reporting and interpretation of study results. For this reason, many journals, and most high-profile ones, require that authors disclose their potential conflicts of interest [80, 84]. In addition, many multi-investigator studies have publication policies that exclude from authorship those with major conflicts of interest.

More extreme is "ghost authorship," where the papers are written by employees of the sponsors, who are not listed as authors, and the academic-based investigators, who may have had little or no role in drafting the manuscript, are given authorship credit. We deplore this practice. We also deplore the practice of listing as authors any who did not truly contribute to the research. In response to these concerns about "guest authorship," many journals now ask for the contribution of each listed author when the manuscript is submitted for publication. (See Chap. 19 for further discussion of these issues.)

References

1. Emanuel EJ, Wendler D, Grady C. What makes clinical research ethical? *JAMA* 2000;283:2701–2711.
2. Emanuel EJ, Grady C, Crouch RA, et al. (eds.). *The Oxford Textbook of Clinical Research Ethics*. Oxford: Oxford University Press, 2008.

3. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research; The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, April 18, 1979. http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.htm.
4. The Nuremberg Code. http://www.hhs.gov/ohrp/references/nurcode.htm.
5. World Medical Association Declaration of Helsinki. http://www.wma.net/e/policy/b3.htm.
6. Council for International Organizations of Medical Sciences (CIOMS). http://www.cioms.ch/frame_guidelines_nov_2002.htm.
7. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N Engl J Med* 1989;321:406–412.
8. Elliott C. Guinea-pigging. *New Yorker.* January 7, 2008.
9. Jonsen AR, Miller FG. Research with healthy volunteers**.** In Emanuel EJ, Grady C, Crouch RA, et al. (eds.). *The Oxford Textbook of Clinical Research Ethics*. Oxford: Oxford University Press, 2008, pp. 481–487.
10. Dickert N, Grady C. Incentives for research particpants. In Emanuel EJ, Grady C, Crouch RA, et al. (eds.). *The Oxford Textbook of Clinical Research Ethics*. Oxford: Oxford University Press, 2008, pp. 386–396.
11. Savulescu J, Spriggs M. The hexamethonium asthma study and the death of a normal volunteer in research. *J Med Ethics* 2002;28:3–4.
12. Suntharalingam G, Perry MR, Ward S, et al. Cytokine storm in a phase 1 trial of the anti-CD28 monoclonal antibody TGN1412. *N Engl J Med* 2006;355:1018–1028.
13. St. Clair EW. The calm after the cytokine storm: lessons from the TGN1412 trial (Commentaries). *J Clin Invest* 2008;118:1344–1347 (correction *J Clin Invest* 2008;118:2365).
14. Agrawal M, Emanuel E. Ethics of phase 1 oncology studies: reexamining the arguments and data. *JAMA* 2003;290:1075–1082.
15. Joffe S, Miller FG. Bench to bedside: mapping the moral terrain of clinical research. *Hastings Cent Rep* 2008;38:30–42.
16. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med* 1987;317:141–145.
17. Shaw LW, Chalmers TC. Ethics in cooperative clinical trials. *Ann NY Acad Sci* 1970; 169:487–495.
18. Byar DP, Simon RM, Friedewald WT, et al. Randomized clinical trials: perspectives on some recent ideas. *N Engl J Med* 1976;295:74–80.
19. Spodick DH. The randomized controlled clinical trial:scientific and ethical basis. *Am J Med* 1982;73:420–425.
20. Royall RM, Bartlett RH, Cornell RG, et al. Ethics and statistics in randomized clinical trials. *Stat Sci* 1991;6:52–88.
21. Dawson L, Zarin DA, Emanuel EJ, et al. Considering usual medical care in clinical trial design. *PLoS Med* 2009;6(9):e1000111. Epub 2009 Sep 29.
22. Holm S, Harris J. The standard of care in multinational research. In Emanuel EJ, Grady C, Crouch RA, et al. (eds.). *The Oxford Textbook of Clinical Research Ethics.* Oxford: Oxford University Press, 2008.
23. Temple RJ, Meyer R. Continued need for placebo in many cases, even when there is effective therapy. *Arch Intern Med* 2003;163:371–373.
24. Freedman B, Weijer C, Glass KC. Placebo orthodoxy in clinical research I: empirical and methodological myths. *J Law Med Ethics* 1996;24:243–251.
25. Freedman B, Glass KC, Weijer C. Placebo orthodoxy in clinical research II: ethical, legal, and regulatory myths. *J Law Med Ethics* 1996;24:252–259.
26. Rothman KJ, Michels KB. The continuing unethical use of placebo controls. *N Engl J Med* 1994;331:394–398.
27. Rothman KJ, Michels KB. Update on unethical use of placebos in randomised trials. *Bioethics* 2003;17:188–204.
28. O'Shea JC, Hafley GE, Greenberg S, et al. Platelet glycoprotein IIb/IIIa integrin blockade with eptifibatide in coronary stent intervention: the ESPRIT trial: a randomized controlled trial. *JAMA* 2001;285:2468–2473.

29. Mann H, London AJ, Mann J. Equipoise in the Enhanced Suppression of the Platelet IIb/IIIa Receptor with Integrilin Trial (ESPRIT): a critical appraisal. *Clin Trials* 2005;2:233–241.
30. Tcheng J. Comment on Mann et al. *Clin Trials* 2005;2:242–243.
31. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use: ICH Harmonised Tripartite Guideline: Choice of Control Group and Related Issues in Clinical Trials E10 (July 2000). http://www.ich.org/cache/compo/276-254-1.html.
32. Miller FG. The ethics of placebo-controlled trials. In Emanuel EJ, Grady C, Crouch RA, et al. (eds.). *The Oxford Textbook of Clinical Research Ethics*. Oxford: Oxford University Press, 2008, pp. 261–272.
33. Levinsky NG. Sounding Board. Nonfinancial conflicts of interest in research. *N Engl J Med* 2002;347:759–761.
34. Weinfurt KP, Hall MA, King NMP, et al. Sounding Board: disclosure of financial relationships to participants in clinical research. *N Engl J Med* 2009;361:916–921.
35. Code of Federal Regulations, Title 45, Part 46. http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.htm.
36. Code of Federal Regulations, Title 21. http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm.
37. Cassileth BR, Zupkis RV, Sutton-Smith K, et al. Informed consent – why are its goals imperfectly realized? *N Engl J Med* 1980;302:896–900.
38. Grunder TM. On the readability of surgical consent forms. *N Engl J Med* 1980;302:900–902.
39. Howard JM, DeMets D, the BHAT Research Group. How informed is informed consent? The BHAT experience. *Control Clin Trials* 1981;2:287–303.
40. Henderson GE, Churchill LR, Davis AM, et al. Clinical trials and medical care: defining the therapeutic misconception. *PLoS Med* 2007; 4:e324.
41. U.S. Food and Drug Administration. Exception from informed consent requirements for emergency research. http://www.fda.gov/RegulatoryInformation/Guidances/ucm127625.htm.
42. Federal Register, volume 61, October 2, 1996, 45 CFR Part 46, pages 5131–5133; Department of Health and Human Services, Waiver of Informed Consent Requirements in Certain Emergency Research. http://www.hhs.gov/ohrp/documents/100296.pdf.
43. Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (amended 2005). http://www.pre.ethics.gc.ca/policy-politique/tcps-eptc/docs/TCPS%20October%202005_E.pdf.
44. European Medicines Agency ICH Topic E6 (R1) Guideline for Good Clinical Practice, January 1997 (corrected July 2002). http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guidline/2009/09/WC500002874.pdf.
45. Bottinger BW, Arntz H-R, Chamberlain DA, et al. Thrombolysis during resuscitation for out-of-hospital cardiac arrest. *N Engl J Med* 2008;359:2651–2662.
46. Burton TM. Despite heart attack deaths, PolyHeme still being tested on trauma patients. *Wall St J* February 22, 2006.
47. Kipnis K, King NMP, Nelson RM. Trials and errors: barriers to oversight of research conducted under the emergency research consent waiver. *IRB* 2006;28:16–19.
48. Karlawish JHT. Sounding Board: Research involving cognitively impaired adults. *N Engl J Med* 2003;348:1389–1392.
49. Rosenstein DL, Miller FG. Research involving those at risk for impaired decision-making capacity. In Emanuel EJ, Grady C, Crouch RA, et al. (eds.). *The Oxford Textbook of Clinical Research Ethics*. Oxford: Oxford University Press, 2008, pp. 437–445.
50. Steinbrook R. How best to ventilate? Trial design and patient safety in studies of the acute respiratory distress syndrome. *N Engl J Med* 2003;348:1393–1401.
51. Silverman HJ, Luce JM, Schwartz J. Protecting subjects with decisional impairment in research: the need for a multifaceted approach. *Am J Respir Crit Care Med* 2004;169:10–14.
52. Love RR, Duc NB, Allred DC, et al. Oophorectomy and tamoxifen adjuvant therapy in premenopausal Vietnamese and Chinese women with operable breast cancer. *J Clin Oncol* 2002;20:2559–2566.

53. Vickers A, Goyal N, Harland R, Rees R. Do certain countries produce only positive results? A systematic review of controlled trials. *Control Clin Trials* 1998;19:159–166.
54. O'Shea JC, Califf RM. International differences in cardiovascular clinical trials. *Am Heart J* 2001;141:875–880.
55. London AL. Responsiveness to host community health needs. In Emanuel EJ, Grady C, Crouch RA, et al. (eds.). *The Oxford Textbook of Clinical Research Ethics*. Oxford: Oxford University Press, 2008, pp. 737–744.
56. Lind SE. Finder's fees for research subjects. *N Engl J Med* 1990; 323:192–195.
57. The Alpha-Tocopherol Beta Carotene Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *N Engl J Med* 1994;330:1029–1035.
58. Miller AB, Buring J, Williams OD. Stopping the carotene and retinal efficacy trial: the viewpoint of the safety and endpoint monitoring committee. In DeMets DL, Furberg CD, Friedman LM. (eds.). *Data Monitoring in Clinical Trials: A Case Studies Approach*. New York: Springer, 2006, pp. 220–227.
59. The Age-Related Eye Disease Study Research Group. Design Paper: The Age-Related Eye Disease Study (AREDS): design implications. AREDS report no. 1. *Control Clin Trials* 1999;20:573–600.
60. Age-Related Eye Disease Study Research Group. A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E, beta carotene, and zinc for age-related macular degeneration and vision loss. AREDS report no. 8. *Arch Ophthalmol* 2001;119:1417–1436 (correction *Arch Ophthalmol* 2008;126:1251).
61. Tegeler CH, Furberg CD. Lessons from warfarin trials in atrial fibrillation: missing the window of opportunity. In DeMets DL, Furberg CD, Friedman LM (eds.). *Data Monitoring in Clinical Trials: A Case Studies Approach*. New York: Springer, 2006, pp. 312–319.
62. Hulley S, Grady D, Bush T, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS). *JAMA* 1998;280:605–613.
63. Rossouw JE, Anderson GL, Prentice RL, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *JAMA* 2002;288:321–333.
64. Hulley SB, Grady D, Vittinghoff E, Williams OD. Consideration of early stopping and other challenges in monitoring the Heart and Estrogen/Progestin Replacement Study. In DeMets DL, Furberg CD, Friedman LM (eds.). *Data Monitoring in Clinical Trials: A Case Studies Approach*. New York: Springer, 2006, pp. 236–247.
65. Wittes J, Barrett-Connor E, Braunwald E, et al. Monitoirng the randomized trials of the Women's Health Initiative: the experience of the Data and Safety Monitoring Board. *Clin Trials* 2007;4:218–234.
66. Peppercorn J, Buss WG, Fost N, Godley PA. The dilemma of data-safety monitoring: provision of significant new data to research participants. *Lancet* 2008;371:527–529.
67. Black HR, Elliott WJ, Grandits G, et al. Principal results of the Controlled Onset Verapamil Investigation of Cardiovascular End Points (CONVINCE) trial. *JAMA* 2003;289:2073–2082.
68. Psaty BM, Rennie D. Stopping medical research to save money: a broken pact with researchers and patients (editorial). *JAMA* 2003;289: 2128–2131.
69. Marcus AD. Paying to keep your drug trial alive. *Wall Street J* April 10, 2007.
70. Health Insurance Portability and Accountability Act (HIPAA). http://privacyruleandresearch.nih.gov/.
71. NIH Data Sharing Policy and Implementation Guidance (updated March 5, 2003). http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm.
72. Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html.
73. Zarin DA, Tse T. Moving toward transparency of clinical trials. *Science* 2008;319: 1340–1342.

74. Wilson D, Meier B. Doctor falsified study on injured G.I.'s, Army says. *The New York Times.* May 12, 2009.
75. Scott J. Withdrawal of a paper (editorial). *J Bone Joint Surg Br* 2009;91:285–286.
76. Fisher B, Anderson S, Redmond CK, et al. Reanalysis and results after 12 years of follow-up in a randomized clinical trial comparing total mastectomy with lumpectomy with or without irradiation in the treatment of breast cancer. *N Engl J Med* 1995;333:1456–1461.
77. Angell M, Kassirer JP. Setting the record straight in the breast cancer trials (editorial). *N Engl J Med* 1994;330:1448–1450.
78. Turner EH, Matthews AM, Linardatos E, et al. Selective publication of antidepressant trials and its influence on apparent efficacy. N *Engl J Med* 2008;358:252–260.
79. Heres S, Davis J, Maino K, et al. Why olanzapine beats risperidone, risperidone beats quetiapine, and quetiapine beats olanzapine: an exploratory analysis of head-to-head comparison studies of second-generation antipsychotics. *Am J Psychiatry* 2006;163:185–194.
80. International Committee of Medical Journal Editors. Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication (Updated October 2008). http://www.icmje.org/.
81. Clinical Trials Registration in ClinicalTrials.gov (Public Law 110-85): Competing Applications and Non-Competing Progress Reports. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-023.html.
82. Federal Register: May 21, 2008 (Volume 73, Number 99). http://edocket.access.gpo.gov/2008/E8-11042.htm.
83. Steinbrook R. Gag clauses in clinical-trial agreements (perspective). *N Engl J Med* 2005;352:2160–2162.
84. Drazen JM, Van Der Weyden MB, Sahni P, et al. Editorial. Uniform format for disclosure of competing interests in ICMJE journals. www.nejm.org October 13, 2009 (10.1056/NEJMe0909052).

Chapter 3
What Is the Question?

The planning of a clinical trial depends on the question that the investigator is addressing. The general objective is usually obvious, but the specific question to be answered by the trial is often not stated well. Stating the question clearly and in advance encourages proper design. It also enhances the credibility of the findings. One would like answers to a number of questions, but the study should be designed with only one major question in mind. This chapter discusses the selection of this primary question and appropriate ways of answering it. In addition, types of secondary and subsidiary questions are reviewed.

The first generation of clinical trials typically compared new interventions to placebo or no treatment on top of best current medical care. They addressed the straight-forward question of whether the new treatment was beneficial, neutral, or harmful compared to placebo or nothing. Since that time, the best medical care has improved dramatically, largely due to the contribution of randomized clinical trials.

Because of this success, new design challenges emerged. Due to the lower event rate in patients receiving best care, the margins for improvement with newer interventions became smaller. This statistical power issue has been addressed in three ways: first, sample sizes have been increased (see Chap. 8); second, there has been an increased reliance on composite outcomes; and third, there is an increased use of surrogate outcomes.

Another consequence was the emergence of trials designed to answer a different type of question. Do alternative treatments that may be equal to, or at least no worse than, existing treatments with regard to the primary outcome convey important benefits in terms of safety, adherence, patient convenience, or cost? These trials are often referred to as noninferiority trials. These trials are discussed later in this chapter and in more detail in Chaps. 5, 8, and 17.

Fundamental Point

Each clinical trial must have a primary question. The primary question, as well as any secondary or subsidiary questions, should be carefully selected, clearly defined, and stated in advance.

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_3, © Springer Science+Business Media, LLC 2010

Selection of the Questions

Primary Question

The primary question should be the one the investigators are most interested in answering and that is capable of being adequately answered. It is the question upon which the sample size of the study is based, and which must be emphasized in the reporting of the trial results. The primary question may be framed in the form of testing a hypothesis because most of the time an intervention is postulated to have a particular outcome which, on the average, will be different from the outcome in a control group [1]. The outcome may be a beneficial clinical event such as improving survival, ameliorating an illness or disease complications, reducing symptoms, or improving quality of life; modifying an intermediate or surrogate characteristic such as blood pressure; or changing a biomarker such as a laboratory value.

Secondary Questions

There may also be a variety of subsidiary or *secondary questions* that are usually related to the primary question. The study may be designed to help address these, or else data collected for the purpose of answering the primary question may also elucidate the secondary questions. They can be of two types. In the first, the response variable is different than that in the primary question. For example, the primary question might ask whether mortality from any cause is altered by the intervention. Secondary questions might relate to incidence of cause-specific death (such as cancer mortality), sex or age-specific mortality, incidence of nonfatal renal failure, or incidence of stroke.

The second type of secondary question relates to *subgroup hypotheses*. For example, in a study of cancer therapy, the investigator may want to look specifically at people by stage of disease at entry into the trial. Such subsets of people in the intervention group can be compared with similar people in the control group. Subgroup hypotheses should be (a) specified before data collection begins, (b) based on reasonable expectations, and (c) limited in number. In any event, the number of participants in any subgroups is usually too small to prove or disprove a subgroup hypothesis. One should not expect significant differences in subgroups unless the trial was specifically designed to detect them. Failure to find significant differences should not be interpreted to mean that they do not exist. Investigators should exercise caution in accepting subgroup results, especially when the overall trial results are not significant. A survey of clinical trialists indicated that inappropriate subgroup analyses were considered as one of the two major sources of distortion of trial findings [2]. Generally, the most useful reason for considering subgroups is to examine consistency of results across predefined subgroups.

There has been recognition that certain subgroups of people have not been adequately represented in clinical research, including clinical trials [3]. In the United States,

this has led to requirements that women and minority populations be included in appropriate numbers in trials [4]. The issue is whether the numbers of participants of each sex and racial/ethnic group must be adequate to answer the key questions that the trial addresses, or whether there must merely be adequate diversity of people. As has been noted, [5, 6] the design of the trial should be driven by reasonable expectations that the intervention will or will not operate materially differently among the various subsets of participants. If so, then it is appropriate to design the trial to detect those differences. If not, adequate diversity with the opportunity to examine subgroup responses at the end of the trial is more appropriate.

Both types of secondary questions raise several methodological issues; for example, if enough statistical tests are done, a few will be significant by chance alone when there is no true intervention effect. An example was provided by the Second International Study of Infarct Survival (ISIS-2), a factorial design trial of aspirin and streptokinase in patients with acute myocardial infarction [7]. Participants born under the Gemini or Libra astrological birth signs had a somewhat greater incidence of vascular and total mortality on aspirin than on no aspirin, whereas for all other signs, and overall, there was an impressive and highly significant benefit from aspirin. Therefore, when a number of tests are carried out, results should be interpreted cautiously. Shedding light or raising new hypotheses is a more proper outcome of these analyses than conclusive answers. See Chap. 17 for further discussion of subgroup analysis.

Both primary and secondary questions should be important and relevant scientifically, medically, or for public health purposes. Participant safety and well-being must always be considered in evaluating importance. Potential benefit and risk of harm should be looked at by the investigator, as well as by local ethical review committees, and often, independent monitoring committees.

Adverse Events

Important questions that can be answered by clinical trials concern adverse events or side effects of therapy (Chap. 12). Here, unlike the primary or secondary questions, it is not always possible to specify in advance the question to be answered. What adverse reactions might occur, and their severity, may be unpredictable. Furthermore, rigorous, convincing demonstration of serious toxicity is usually not achieved because it is generally thought unethical to continue a study to the point at which a drug has been conclusively shown to be more harmful than beneficial [8–10]. Investigators traditionally monitor a variety of laboratory and clinical measurements, look for possible adverse effects, and compare these in the intervention and control groups. Statistical significance and the previously mentioned problem of multiple response variables become secondary to clinical judgment and participant safety. While this will lead to the conclusion that some purely chance findings are labeled as adverse effects, moral responsibility to the participants requires a conservative attitude toward safety monitoring, particularly if an alternative therapy is available.

Ancillary Questions, Substudies

Often a clinical trial can be used to answer questions which do not bear directly on the intervention being tested, but which are nevertheless of interest. The structure of the trial and the ready access to participants may make it the ideal vehicle for such investigations. Weinblatt, Ruberman, and colleagues reported that a low level of education among survivors of a myocardial infarction was a marker of poor risk of survival [11]. The authors subsequently evaluated whether the educational level was an indicator of psychosocial stress [12]. To further investigate these findings, they performed a study ancillary to the Beta-Blocker Heart Attack Trial (BHAT), [13] a trial which evaluated whether the regular administration of propranolol could reduce 3-year mortality in people with acute myocardial infarctions. Interviews assessing factors such as social interaction, attitudes, and personality were conducted in over 2,300 men in the ancillary study [14]. Inability to cope with high life stress and social isolation were found to be significantly and independently associated with mortality. Effects of low education were accounted for by these two factors. By enabling the investigators to perform this study, the BHAT provided an opportunity to examine an important issue in a large sample, even though it was peripheral to the main question.

In the Studies of Left Ventricular Dysfunction (SOLVD), [15] the investigators evaluated whether an angiotensin converting enzyme inhibitor would reduce mortality in symptomatic and asymptomatic subjects with impaired cardiac function. In selected participants, special studies were done with the objective of getting a better understanding of the disease process and of the mechanisms of action of the intervention. These substudies did not require the large sample size of the main studies (over 6,000 participants). Therefore, most participants in the main trial had a relatively simple and short evaluation and did not undergo the expensive and time-consuming procedures or interviews demanded by the substudies. This combination of a rather limited assessment in many participants, designed to address an easily monitored response variable, and detailed measurements in subsets, can be extremely effective.

Natural History

Though it is not intervention-related, a sometimes valuable use of the collected data, especially in long-term trials, is a natural history study in the control group, if it consists of placebo or no treatment. This was done in some early cardiovascular clinical trials [16]. Early AIDS trials yielded considerable information about natural history of the disease at a time when there was considerable uncertainty. If the control group is either on placebo or no systematic treatment, various baseline factors may be studied for their relation to specific outcomes. Assessment of the prognostic importance of these factors can lead to better understanding of the disease under study and development of new hypotheses. Of course, generally only predictive association – and not necessarily causation – may properly be inferred from such data. The study participants may be a highly selected group of people who, although

they may be on placebo, are receiving various concomitant therapies for their condition. In addition, in some fields such as cardiology and oncology, large, well-characterized observational cohorts exist, lessening the need for clinical trial control groups to provide natural history data. In selected conditions or circumstances, however, these kinds of control group analyses may be important, as long as the findings are interpreted in the context of selection of the study sample.

Since they are not trial hypotheses, specific natural history questions need not be specified in advance. However, properly designed baseline forms require some advance consideration of which factors might be related to outcome. After the study has started, going back to ascertain missing baseline information in order to answer natural history questions is generally a fruitless pursuit. At the same time, collecting large amounts of baseline data on the slight chance that they might provide useful information costs money, consumes valuable time, and may lead to less careful collection of important data. It is better to restrict data collection to those baseline factors that are known, or seriously thought, to have an impact on prognosis.

Large, Simple Clinical Trials

As discussed in more detail in Chap. 5, the concept of "large, simple clinical trials" has become popular [17]. The general idea is that for common conditions and important outcomes such as total mortality, even modest benefits of intervention, particularly interventions that are easily implemented in a large population are important. Because an intervention is likely to have similar effects in different sorts of participants, careful characterization of people at entry may be unnecessary. The study must have unbiased allocation of participants to intervention or control and unbiased assessment of outcome. Sufficiently large numbers of participants are more important in providing the power necessary to answer the question than careful attention to quality of data. This model depends upon a relatively easily administered intervention, brief forms, and an easily ascertained outcome, such as a fatal or nonfatal event.

Superiority vs. Noninferiority Trials

As mentioned in the introduction to this chapter, traditionally, most trials were designed to establish whether a new intervention on top of usual or standard care was superior to that care alone (or that care plus placebo). If there were no effective treatments, the new intervention was compared to just placebo. As discussed in Chap. 8, these trials are generally two-sided. That is, the trial is designed to see whether the new intervention is better or worse than the control.

Once effective therapies were developed, more trials were designed to demonstrate that a new intervention is not worse than the control by some prespecified amount. As noted earlier, the motivation for such a question is that the new intervention might not be better than standard treatment on the primary or important secondary

outcomes, but may be less toxic, more convenient, less invasive, or have some other attractive feature. The challenge is to define what is meant by "not worse than." This will be referred to as the "margin of indifference," or δ, meaning that if the new intervention is not less effective than this margin, its use might be of interest given the other features. In the analysis of this design, the 95% upper confidence limit would need to be less than this margin in order to claim noninferiority. Defining δ is challenging and will be discussed in Chap. 5.

The question in a noninferiority trial is different than in a superiority trial and affects both the design and conduct of the trial. For example, in the superiority trial, poor adherence will lead to a decreased ability, or power, to detect a meaningful difference. For a noninferiority trial, poor adherence will diminish important differences and bias the results toward a noninferiority claim. Thus, great care must be taken in defining the question, the sensitivity of the outcome measures to the intervention being evaluated, and the adherence to the intervention during the conduct of the trial.

Intervention

When the question is conceived, investigators, at the very least have in mind a class or type of intervention. More commonly, they know the precise drug, procedure, or lifestyle modification they wish to study. In reaching such a decision, they need to consider several aspects. First, the potential benefit of the intervention must be maximized while possible toxicity is kept to a minimum. Thus, dose of drug or intensity of rehabilitation and frequency and route of administration are key factors that need to be determined. Can the intervention be standardized, and remain reasonably stable over the duration of the trial? Investigators must also decide whether to use a single drug, biologic, or device, fixed or adjustable doses of drugs, sequential drugs, or drug or device combinations. Devices in particular undergo frequent modifications and updates. Investigators need to be satisfied that newer versions that appear during the course of the trial function sufficiently similarly in important ways to the older versions so that combining data from the versions is appropriate. Of course, an investigator can use only the older version (if it is still available), but the trial will then be criticized for employing the outdated version. In gene transfer studies, the nature of the vector, as well as the actual gene, may materially affect the outcome, particularly when it comes to adverse effects.

Not only the nature of the intervention, but what constitutes the control group regimen must also be considered for ethical reasons, as discussed in Chap. 2, and study design reasons, as discussed in Chap. 5.

Second, the availability of the drug or device for testing needs to be determined. If it is not yet licensed, special approval from the regulatory agency and cooperation or support by the manufacturer are required.

Third, investigators must take into account design aspects, such as time of initiation and duration of the intervention, need for special tests or laboratory facilities, and the logistics of blinding in the case of drug studies. Certain kinds of interventions, such as surgical

procedures, device implantation, vaccines, and gene transfer may have long-term or even life-long effects. Therefore, investigators might need to incorporate plans for long-term assessment. There had been reports that drug-eluting stents, used in percutaneous coronary intervention, perhaps had a greater likelihood of restenosis than bare-metal stents [18, 19]. Follow-up studies seemed to assuage these concerns [20]. Nevertheless, investigators must consider incorporating plans for long-term assessment.

Response Variables

Response variables are outcomes measured during the course of the trial, and they define and answer the questions. A response variable may be total mortality, death from a specific cause, incidence of a disease, a complication or specific adverse effect of a disease, symptomatic relief, a clinical finding, a laboratory measurement, or the cost and ease of administering the intervention. If the primary question concerns total mortality, the occurrence of deaths in the trial clearly answers the question. If the primary question involves severity of arthritis, on the other hand, extent of mobility or a measure of freedom from pain may be reasonably good indicators. In other circumstances, a specific response variable may only partially reflect the overall question. As seen from the above examples, the response variable may show a change from one discrete state (living) to another (dead), from one discrete state to any of several other states (changing from one stage of disease to another) or from one level of a continuous variable to another. If the question can be appropriately defined using a continuous variable, the required sample size may be reduced (Chap. 8). However, the investigator needs to be careful that this variable and any observed differences are clinically meaningful and relevant and that the use of a continuous variable is not simply a device to reduce sample size.

In general, a single response variable should be identified to answer the primary question. If more than one are used, the probability of getting a nominally significant result by chance alone is increased (Chap. 17). In addition, if several response variables give inconsistent results, interpretation becomes difficult. The investigator would then need to consider which outcome is most important, and explain why the others gave conflicting results. Unless she has made the determination of relative importance prior to data collection, her explanations are likely to be unconvincing.

Although the practice is not advocated, there may be circumstances when more than one "primary" response variable needs to be looked at. This may be the case when an investigator truly cannot state which of several response variables relates most closely to the primary question. Ideally, the trial would be postponed until this decision can be made. However, overriding concerns, such as increasing use of the intervention in general medical practice, may compel her to conduct the study earlier. In these circumstances, rather than arbitrarily selecting one response variable which may, in retrospect, turn out to be inappropriate, investigators prefer to list several "primary" outcomes. For instance, in the Urokinase Pulmonary Embolism Trial [21] lung scan, arteriogram and hemodynamic measures were given as the "primary" response variables in assessing the effectiveness of the agents urokinase and streptokinase.

Chapter 8 discusses the calculation of sample size when a study with several primary response variables is designed.

Combining events to make up a response variable might be useful if any one event occurs too infrequently for the investigator reasonably to expect a significant difference without using a large number of participants. In answering a question where the response variable involves a combination of events, only *one event per participant* should be counted. That is, the analysis is by participant, not by event.

One kind of combination response variable involves two kinds of events. This has been termed a *composite outcome*. It must be emphasized, however, that the combined events should be capable of meaningful interpretation such as being related through a common underlying condition or responding to the same presumed mechanism of action of the agent. In a study of heart disease, combined events might be death from coronary heart disease plus nonfatal myocardial infarction. This is clinically meaningful since death from coronary heart disease and nonfatal myocardial infarction might together represent a measure of serious coronary heart disease. Difficulties in interpretation can arise if the results of each of the components in such a response variable are inconsistent. In the Physicians' Health Study report of aspirin to prevent cardiovascular disease, there was no difference in mortality, a large reduction in myocardial infarction, and an increase in stroke, primarily hemorrhagic [22]. In this case, cardiovascular mortality was the primary response variable, rather than a combination. If it had been a combination, the interpretation of the results would have been even more difficult than it was [23]. Even more troublesome is the situation where one of the components in the combination response variable is far less serious than the others. For example, if occurrence of angina pectoris or a revascularization procedure is added, as is commonly done, interpretation can be problematic. Not only are these less serious than cardiovascular death or myocardial infarction, they often occur more frequently. Thus, if overall differences between groups are seen, are these results driven primarily by the less serious components? What if the results for the more serious components (e.g., death) trend in the opposite directions? This is not just theoretical. For example, the largest difference between intervention and control in the Myocardial Ischemia Reduction with Aggressive Cholesterol Lowering (MIRACL) trial was seen in the least serious of the four components; the one that occurred most often in the control group [24]. A survey of published trials in cardiovascular disease that used composite response variables showed that half had major differences in both importance and effect sizes of the individual components [25]. Those components considered to be most important had, on average, smaller benefits than the more minor ones. See Chap. 17 for a discussion of analytic and interpretation issues if the components of the composite outcome go in different directions or have other considerable differences in the effect size.

When this kind of combination response variable is used, the rules for interpreting the results and for possibly making regulatory claims about individual components should be established in advance. This is particularly important if there are major differences in seriousness. A survey of the cardiovascular literature found that the use of composite outcomes (often with three or four components) was common, and the components varied in importance [26]. One possible approach is to require that the most serious individual components show the same trend as the overall result.

Some have suggested giving each component weights, depending on the seriousness [27, 28]. Although it has sample size implications, it is probably preferable to include in the combined primary response variable only those components that are truly serious and to assess the other components as secondary outcomes.

Another kind of combination response variable involves multiple events of the same sort. Rather than simply asking whether an event has occurred, the investigator can look at the frequency with which it occurs. This may be a more meaningful way of looking at the question than seeking a yes–no outcome. For example, frequency of recurrent transient ischemic attacks or epileptic seizures within a specific follow-up period might comprise the primary response variable of interest. Simply adding up the number of recurrent episodes and dividing by the number of participants in each group in order to arrive at an average is improper. Multiple events in an individual are not independent, and averaging gives undue weight to those with more than one episode. One approach is to compare the number of participants with none, one, two, or more episodes; that is, the distribution, by individual, of the number of episodes.

Sometimes, study participants enter a trial with a condition that is exhibited frequently. For example, they may have had several episodes of atrial fibrillation in the previous weeks or may drink alcohol to excess several days a month. Trial eligibility criteria may even require a minimum number of such episodes. A trial of a new treatment for alcohol abuse may require participants to have at least six alcoholic drinks a day for at least 7 days over the previous month. The investigator needs to decide what constitutes a beneficial response. Is it complete cessation of drinking? Reducing the number of drinks to some fixed level (e.g., no more than two on any given day)? Reducing alcohol intake by some percent, and if so, what percent? Does this fixed level or percent differ depending on the intake at the start of the trial? Decisions must be made based on knowledge of the disease or condition, the kind of intervention, and the expectations of how the intervention will work. The clinical importance of improvement versus “cure” must also be considered.

Specifying the Question

Regardless of whether an investigator is measuring a primary or secondary response variable, certain rules apply. First, she should define and write the questions in advance, being as specific as possible. She should not simply ask, “Is *A* better than *B*?” Rather, she should ask, “In population *W* is drug *A* at daily dose *X* more efficacious in improving *Z* by *Q* amount over a period of time *T* than drug *B* at daily dose *Y*?” Implicit here is the magnitude of the difference that the investigator is interested in detecting. Stating the questions and response variables in advance is essential for planning of study design and calculation of sample size. As shown in Chap. 8, sample size calculation requires specification of the response variables as well as estimates of the effect of intervention. In addition, the investigator is forced to consider what she means by a successful intervention. For example, does the intervention need to reduce mortality by 10% or 25% before a recommendation for its general use is made? Since such recommendations

also depend on the frequency and severity of adverse effects, a successful result cannot be completely defined beforehand. However, if a 10% reduction in mortality is clinically important, that should be stated, since it has sample size implications. Specifying response variables and anticipated benefit in advance also eliminates the possibility of the legitimate criticism that can be made if the investigator looked at the data until she found a statistically significant result and then decided that *that* response variable was what she really had in mind all the time. Investigators have changed the primary response variable partway through a trial because of concerns about adequate power to answer the original question [29, 30]. On occasion, however, the reported primary response variable was changed without clear rationale and after the data had been examined [31, 32].

Second, the primary response variable must be capable of being assessed in all participants. Selecting one response variable to answer the primary question in some participants, and another response variable to answer the same primary question in other participants is not a legitimate practice. It implies that each response variable answers the question of interest with the same precision and accuracy; i.e., that each measures exactly the same thing. Such agreement is unlikely. Similarly, response variables should be measured in the same way for all participants. Measuring a given variable by different instruments or techniques implies that the instruments or techniques yield precisely the same information. This rarely, if ever, occurs. If response variables can be measured only one way in some participants and another way in other participants, two separate studies are actually being performed, each of which is likely to be too small.

Third, unless there is a combination primary response variable in which the participant remains at risk of having additional events, participation generally ends when the primary response variable occurs. "Generally" is used here because, unless death is the primary response variable, the investigator may well be interested in certain events, including adverse events, subsequent to the occurrence of the primary response variable. These events will not change the analysis of the primary response variable but may affect the interpretation of results. For example, deaths taking place after a nonfatal primary response variable has already occurred, but before the official end of the trial as a whole, may be of interest. On the other hand, if a secondary response variable occurs, the participant should remain in the study (unless, of course, it is a fatal secondary response variable). He must continue to be followed because he is still at risk of developing the primary response variable. A study of heart disease may have, as its primary question, death from coronary heart disease and, as a secondary question, incidence of nonfatal myocardial infarction. If a participant suffers a nonfatal myocardial infarction, this counts toward the secondary response variable. However, he ought to remain in the study for analytic purposes and be at risk of dying (the primary response variable) and of having other adverse events. This is true whether or not he is continued on the intervention regimen. If he does not remain in the study for purposes of analysis of the primary response variable, bias may result. (See Chap. 17 for further discussion of participant withdrawal.)

Fourth, response variables should be capable of unbiased assessment. Truly double-blind studies have a distinct advantage over other studies in this regard.

If a trial is not double-blind (Chap. 7), then, whenever possible, response variable assessment should be done by people who are not involved in participant follow-up and who are blinded to the identity of the study group of each participant. Independent reviewers are often helpful. Of course, the use of blinded or independent reviewers does not solve the problem of bias. Unblinded investigators sometimes fill out forms and the participants may be influenced by the investigators. This may be the case during an exercise performance test, where the impact of the person administering the test on the results may be considerable. Some studies arrange to have the intervention administered by one investigator and response variables evaluated by another. Unless the participant is blinded to his group assignment (or otherwise unable to communicate), this procedure is also vulnerable to bias. One solution to this dilemma is to use only "hard," or objective, response variables (which are unambiguous and not open to interpretation, such as total mortality or various imaging or laboratory measures). This assumes complete and honest ascertainment of outcome. Double-blind studies have the advantage of allowing the use of softer response variables, since the risk of assessment bias is minimized.

Fifth, it is important to have response variables that can be ascertained as completely as possible. A hazard of long-term studies is that participants may fail to return for follow-up appointments. If the response variable is one that depends on an interview or an examination, and participants fail to return for follow-up appointments information will be lost. Not only will it be lost, but it may be differentially lost in the intervention and control groups. Death or hospitalizations are useful response variables because the investigator can usually ascertain vital status or occurrence of a hospital admission, even if the participant is no longer active in a study. However, only in a minority of clinical trials are they appropriate.

Sometimes, participants withdraw their consent to be in the trial after the trial has begun. In such cases, the investigator should ascertain whether the participant is simply refusing to return for follow-up visits but is willing to have his data used, including data that might be obtained from public records; is willing to have only data collected up to the time of withdrawal used in analyses; or is asking that all of his data be deleted from the study records.

All clinical trials are compromises between the ideal and the practical. This is true in the selection of primary response variables. The most objective or those most easily measured may occur too infrequently, may fail to define adequately the primary question, or may be too costly. To select a response variable which can be reasonably and reliably assessed and yet which can provide an answer to the primary question requires judgment. If such a response variable cannot be found, the wisdom of conducting the trial should be re-evaluated.

Biomarkers and Surrogate Response Variables

A common criticism of clinical trials is that they are expensive and of long duration. This is particularly true for trials which use the occurrence of clinical events as the

primary response variable. It has been suggested that response variables which are continuous in nature might substitute for the clinical outcomes. Thus, instead of monitoring cardiovascular mortality or myocardial infarction, an investigator could examine progress of atherosclerosis by means of angiography or ultrasound imaging, or change in cardiac arrhythmia by means of ambulatory electrocardiograms or programmed electrical stimulation. In the cancer field, change in tumor size might replace mortality. In AIDS trials, change in CD-4 lymphocyte level has been used as a response to treatment instead of incidence of AIDS in HIV positive patients or mortality. Improved bone mineral density has been used as a surrogate for reduction in fractures.

An argument for use of these "surrogate response variables" is that since the variables are continuous, the sample size can be smaller and the study less expensive than otherwise. Also, changes in the variables are likely to occur before the clinical event, shortening the time required for the trial. Wittes et al. [33] discuss examples of savings in sample size by the use of surrogate response variables.

It has been argued that in the case of truly life-threatening diseases (e.g., AIDS in its early days, certain cancers), clinical trials should not be necessary to license a drug or other intervention. Given the severity of the condition, lesser standards of proof should be required. If clinical trials are done, surrogate response variables ought to be acceptable, as speed in determining possible benefit is crucial. Potential errors in declaring an intervention useful may therefore not be as important as early discovery of a truly effective treatment.

Even in such instances, however, one should not uncritically use surrogate endpoints [34, 35]. It was known for years that the presence of ventricular arrhythmias correlated with increased likelihood of sudden death and total mortality in people with heart disease, [36] as it was presumably one mechanism for the increased mortality. Therefore, it was common practice to administer antiarrhythmic drugs with the aim of reducing the incidence of sudden cardiac death [37, 38]. The Cardiac Arrhythmia Suppression Trial recently demonstrated, however, that drugs which effectively treated ventricular arrhythmias were not only ineffective in reducing sudden cardiac death, but actually caused increased mortality [39, 40].

A second example concerns the use of inotropic agents in people with heart failure. These drugs had been shown to improve exercise tolerance and other symptomatic manifestations of heart failure [41]. It was expected that mortality would also be reduced. Unfortunately, clinical trials subsequently showed that mortality was worsened [42, 43].

Another example from the cardiovascular field is the Investigation of Lipid Level Management to Understand its Impact in Atherosclerotic Events (ILLUMINATE) trial. In this trial, the combination of torcetrapib and atorvastatin was compared with atorvastatin alone in people with cardiovascular disease or diabetes. Despite the expected impressive and highly statistically significant increase in HDL-cholesterol and decrease in LDL-cholesterol in the combination group, there was an increase in all-cause mortality and major cardiovascular events [44]. Thus, even though it is well known that lowering LDL-cholesterol (and possibly increasing HDL-cholesterol) can lead to a reduction in coronary heart disease events, some interventions might have unforeseen adverse consequences.

It was noted that the level of CD-4 lymphocytes in the blood is associated with severity of AIDS. Therefore, despite some concerns [45], a number of clinical trials used change in CD-4 lymphocyte concentration as an indicator of disease status. If the level rose, the drug was considered to be beneficial. Lin et al., however, argued that CD-4 lymphocyte count accounts for only part of the relationship between treatment with zidovudine and outcome [46]. Choi et al. came to similar conclusions [47]. In a trial comparing zidovudine with zalcitabine, zalcitabine was found to lead to a slower decline in CD-4 lymphocytes but had no effect on the death rate [48]. Also troubling were the results of a large trial which, although showing an early rise in CD-4 lymphocytes, did not demonstrate any long-term benefit from zidovudine [49]. Whether zidovudine or another treatment was, or was not, truly beneficial is not the issue here. The main point is that the effect of a drug on a surrogate endpoint (CD-4 lymphocytes) is not always a good indicator of clinical outcome. This is summarized by Fleming, who notes that the CD-4 lymphocyte count showed positive results in seven out of eight trials where clinical outcomes were also positive. However, the CD-4 count was also positive in six out of eight trials in which the clinical outcomes were negative [35].

Similar seemingly contradictory results have been seen with cancer clinical trials. In trials of 5-fluorouracil plus leucovorin compared with 5-fluorouracil alone, the combination led to significantly better tumor response, but no difference in survival [50]. Fleming cites other cancer examples as well [35]. Sodium fluoride, because of its stimulation of bone formation, was widely used in the treatment of osteoporosis. Despite this, it was found in a trial in women with postmenopausal osteoporosis to increase skeletal fragility [51].

These examples do not mean that surrogate response variables should never be used in clinical trials. Nevertheless, they do point out that they should only be used after considering the advantages and disadvantages, recognizing that erroneous conclusions about interventions might occasionally be reached.

Prentice has summarized two key criteria that must be met if a surrogate response variable is to be useful [52]. First, the surrogate must correlate with the true clinical outcome, which most proposed surrogates would likely do. Second, for a surrogate to be valid, it must capture the full effect of the intervention. For example, a drug might lower blood pressure or serum LDL-cholesterol, but as in the ILLUMINATE trial example, have some other deleterious effect that would negate any benefit or even prove harmful.

Another factor is whether the surrogate variable can be assessed accurately and reliably. Is there so much measurement error that, in fact, the sample size requirement increases or the results are questioned? Additionally, will the evaluation be so unacceptable to the participant that the study will become infeasible? If it requires invasive techniques, participants may refuse to join the trial, or worse, discontinue participation before the end. Measurement can require expensive equipment and highly trained staff, which may, in the end, make the trial more costly than if clinical events are monitored. The small sample size of surrogate response variable trials may mean that important data on safety are not obtained [53]. Finally, will the conclusions of the trial be accepted by the scientific and medical communities? If there

is insufficient acceptance that the surrogate variable reflects clinical outcome, in spite of the investigator's conviction, there is little point in using such variables.

Many drugs have been approved by regulatory agencies on the basis of surrogate response variables. We think that, except in rare instances, whenever interventions are approved by regulatory bodies on the basis of surrogate response variables, further clinical studies should be conducted afterward. In all decisions regarding approval, the issues of biologic plausibility, risk, benefits, and history of success must be considered.

When are surrogate response variables useful? The situation of extremely serious conditions has been mentioned. Other than that, surrogate response variables are useful in early phase development studies, as an aid in deciding on proper dosage and whether the anticipated biologic effects are being achieved. They can help in deciding whether, and how best, to conduct the late phase trials which almost always should employ clinical response variables.

General Comments

Although this text attempts to provide straightforward concepts concerning the selection of study response variables, things are rarely as simple as one would like them to be. Investigators often encounter problems related to design, data monitoring and ethical issues, and interpretation of study results.

In long-term studies of participants at high-risk, where total mortality is not the primary response variable, many may nevertheless die. They are, therefore, removed from the population at risk of developing the response variable of interest. Even in relatively short studies, if the participants are seriously ill, death may occur. In designing studies, therefore, if the primary response variable is a continuous measurement, a nonfatal event, or cause-specific mortality, the investigator needs to consider the impact of total mortality for two reasons. First, it will reduce the effective sample size. One would like to allow for this reduction by estimating the overall mortality and increasing sample size accordingly. However, a methodology for estimating mortality and increasing sample size is not yet well defined. Second, if mortality is related to the intervention, either favorably or unfavorably, excluding from study analysis those who die may bias results for the primary response variable.

One solution, whenever the risk of mortality is high, is to choose total mortality as the primary response variable. Alternatively, the investigator can combine total mortality with a pertinent nonfatal event as a combined primary response variable. Neither of these solutions may be appropriate and, in that case, the investigator should monitor total mortality as well as the primary response variable. Evaluation of the primary response variable will then need to consider those who died during the study, or else the censoring may bias the comparison.

Investigators need to monitor total mortality-as well as any other adverse occurrence-during a study, regardless of whether or not it is the primary response variable (see Chap. 16). The ethics of continuing a study which, despite a favorable trend for

the primary response variable, shows equivocal or even negative results for secondary response variables, or the presence of major adverse effects, are questionable. Deciding what to do is difficult if an intervention is giving promising results with regard to death from a specific cause (which may be the primary response variable), yet total mortality is unchanged or increased. An independent monitoring committee has proved extremely valuable in such circumstances (Chap. 16).

Finally, conclusions from data are not always clear-cut. Issues such as alterations in quality of life or annoying long-term side effects may cloud results that are clear with regard to primary response variables such as increased survival. In such circumstances, the investigator must offer her best assessment of the results but should report sufficient detail about the study to permit others to reach their own conclusions (Chap. 19).

References

1. Cutler SJ, Greenhouse SW, Cornfield J, Schneiderman MA. The role of hypothesis testing in clinical trials: biometrics seminar. *J Chronic Dis* 1966;19:857–882.
2. Al-Marzouki S, Roberts I, Marshall T, Evans S. The effect of scientific misconduct on the results of clinical trials: a Delphi survey. *Contemp Clin Trials* 2005;26:331–337.
3. Angell M. Caring for women's health – what is the problem? (editorial). *N Engl J Med* 1993;329:271–272.
4. NIH Revitalization Act, Subtitle B, Part 1, Sec. 131–133, June 10, 1993.
5. Freedman LS, Simon R, Foulkes MA, et al. Inclusion of women and minorities in clinical trials and the NIH Revitalization Act of 1993 – the perspective of NIH clinical trialists. *Control Clin Trials* 1995;16:277–285.
6. Piantadosi S, Wittes J. Letter to the editor. *Control Clin Trials* 1993;14:562–567.
7. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet* 1988; *ii* 349-360.
8. Shimkin MB. The problem of experimentation on human beings. I. The research worker's point of view. *Science* 1953;117:205–207.
9. Chalmers TC. Invited remarks: national conference on clinical trials methodology. *Clin Pharmacol Ther* 1979;25:649–650.
10. Stamler J. Invited remarks: national conference on clinical trials methodology. *Clin Pharmacol Ther* 1979;25:651–654.
11. Weinblatt E, Ruberman W, Goldberg JD, et al. Relation of education to sudden death after myocardial infarction. *N Engl J Med* 1978;299:60–65.
12. Ruberman W, Weinblatt E, Goldberg JD, Chaudhary BS. Education, psychosocial stress, and sudden cardiac death. *J Chronic Dis* 1983;36:151–160.
13. Beta-Blocker Heart Attack Trial Research Group. A randomized trial of propranolol in patients with acute myocardial infarction. 1. Mortality results. *JAMA* 1982;247:1707–1714.
14. Ruberman W, Weinblatt E, Goldberg JD, Chaudhary BS. Psychosocial influences on mortality after myocardial infarction. *N Engl J Med* 1984;311:552–559.
15. The SOLVD Investigators. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *N Engl J Med* 1991;325: 293–302.
16. Schlant RC, Forman S, Stamler J, Canner PL. The natural history of coronary heart disease: prognostic factors after recovery from myocardial infarction in 2,789 men. The 5-year findings of the Coronary Drug Project. *Circulation* 1982;66:401–414.
17. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med* 1984;3:409–422.

18. McFadden E, Stabile E, Regar E, et al. Late thrombosis in drug-eluting coronary stents after discontinuation of antiplatelet therapy. *Lancet* 2004;364:1519–1521.
19. Ong AT, McFadden EP, Regar E, et al. Late angiographic stent thrombosis (LAST) events with drug eluting stents. *J Am Coll Cardiol* 2005;45:2088–2092.
20. Mauri L, Hsieh W-H, Massaro JM, et al. Stent thrombosis in randomized clinical trials of drug-eluting stents. *N Engl J Med* 2007;356:1020–1029.
21. Urokinase Pulmonary Embolism Trial Study Group. Urokinase pulmonary embolism trial: phase I results. *JAMA* 1970;214:2163–2172.
22. Steering Committee of the Physicians' Health Study Research Group. Final report on the aspirin component of the ongoing Physicians' Health Study. *N Engl J Med* 1989;321:129–135.
23. Cairns J, Cohen L, Colton T, et al. Data Monitoring Board of the Physicians' Health Study. Issues in the early termination of the aspirin component of the Physicians' Health Study. *Ann Epidemiol* 1991;1:395–405.
24. Schwartz GG, Olsson AG, Ezekowitz MD, et al., for the Myocardial Ischemia Reduction with Aggressive Cholesterol Lowering (MIRACL) Study Investigators. Effects of atorvastatin on early ischemic events in acute coronary syndromes: the MIRACL study: a randomized controlled trial. *JAMA* 2001;285:1711–1718.
25. Ferreira-Gonzalez I, Busse JW, Heels-Ansdell D, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *Br Med J* 2007; 334:756–757.
26. Lim E, Brown A, Helmy A, et al. Composite outcomes in cardiovascular research: a survey of randomized trials. *Ann Intern Med* 2008;149:612–617.
27. Neaton JD, Wentworth DN, Rhame F, et al. Methods of studying intervention: considerations in choice of a clinical endpoint for AIDS clinical trials. *Stat Med* 1994;13:2107–2125.
28. Hallstrom AP, Litivin PE, Weaver WD. A method of assigning scores to the components of a composite outcome: an example from the MITI trial. *Control Clin Trials* 1992;13:148–155.
29. Julian D. The data monitoring experience in the Carvedilol Post-Infarct Survival Control in Left Ventricular Dysfunction Study: hazards of changing primary outcomes. In: DeMets DL, Furberg CD, Friedman LM (eds) *Data Monitoring in Clinical Trials: A Case Studies Approach*. New York: Springer, 2006.
30. The PEACE Trial Investigators. Angiotensin-converting enzyme inhibition in stable coronary artery disease. *N Engl J Med* 2004;351:2058–2068.
31. The Anturane Reinfarction Trial Research Group. Sulfinpyrazone in the prevention of sudden death after myocardial infarction. *N Engl J Med* 1980;302:250–256.
32. Anturane Reinfarction Trial Policy Committee. The Anturane Reinfarction Trial: reevaluation of outcome. *N Engl J Med* 1982;306:1005–1008.
33. Wittes J, Lakatos E, Probstfield J. Surrogate endpoints in clinical trials: cardiovascular diseases. *Stat Med* 1989; 8:415–425.
34. DeMets DL, Fleming T. Surrogate endpoints in clinical trials: are we being misled? *Ann Intern Med* 1996;125:605–613.
35. Fleming TR. Surrogate markers in AIDS and cancer trials. *Stat Med* 1994;13:1423–1435.
36. Bigger JT Jr, Fleiss JL, Kleiger R, et al., The Multicenter Post-Infarction Research Group. The relationships among ventricular arrhythmias, left ventricular dysfunction, and mortality in the 2 years after myocardial infarction. *Circulation* 1984;69:250–258.
37. Vlay SC. How the university cardiologist treats ventricular premature beats: a nationwide survey of 65 university medical centers. *Am Heart J* 1985;110:904–912.
38. Morganroth J, Bigger JT Jr, Anderson JL. Treatment of ventricular arrhythmias by United States cardiologists: a survey before the Cardiac Arrhythmia Suppression Trial (CAST) results were available. *Am J Cardiol* 1990;65:40–48.
39. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N Engl J Med* 1989;321:406–412.
40. The Cardiac Arrhythmia Suppression Trial II Investigators. Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. *N Engl J Med* 1992;327:227–233.

41. Packer M. Vasodilator and inotopic drugs for the treatment of chronic heart failure: distinguishing hype from hope. *J Am Coll Cardiol.* 1988;12:1299–1317.
42. Packer M, Carver JR, Rodehoffer RT, et al. Effect of oral milrinone on mortality in severe chronic heart failure. *N Engl J Med* 1991;325:1468–1475.
43. The Xamoterol in Severe Heart Failure Study Group. Xamoterol in severe heart failure. *Lancet* 1990;336:1–6; correction *Lancet* 1990;336:698.
44. Barter PJ, Caulfield M, Eriksson M, et al., for the ILLUMINATE Investigators. Effects of torcetrapib in patients at high risk for coronary events. *N Engl J Med* 2007;357:2109–2122.
45. Cohen J. Searching for markers on the AIDS trail. *Science* 1992;258:388–390.
46. Lin DY, Fischl MA, Schoenfeld DA. Evaluating the role of CD-4 lymphocyte counts as surrogate endpoints in human immunodeficiency virus clinical trials. *Stat Med* 1993;12:835–842.
47. Choi S, Lagakos SW, Schooley RT, Volberding PA. CD4+ lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV infection taking zidovidine. *Ann Intern Med* 1993;118:674–680.
48. Fischl MA, Olson RM, Follansbee SE, et al. Zalcitabine compared with zidovudine in patients with advanced HIV-1 infection who received previous zidovudine therapy. *Ann Intern Med* 1993;118:762–769.
49. Aboulker JP, Swart AM. Preliminary analysis of the Concorde trial. *Lancet* 1993;341:889–890.
50. Advanced Colorectal Cancer Meta-Analysis Project. Modulation of fluoruracil by leucovorin in patients with advanced colorectal cancer. Evidence in terms of response rate. *J Clin Oncol* 1992;10:896–903.
51. Riggs BL, Hodgson SF, O'Fallon WM, et al. Effect of fluoride treatment on the fracture rate in postmenopausal women with osteoporosis. *N Engl J Med* 1990;322:802–809.
52. Prentice RL. Surrogate endpoints in clinical trials: definitions and operational criteria. *Stat Med* 1989;8:431–440.
53. Ray WA, Griffin MR, Avorn J. Sounding board: evaluating drugs after their approval for clinical use. *N Engl J Med* 1993;329:2029–2032.

Chapter 4
Study Population

Defining the study population is an integral part of posing the primary question. It is not enough to claim that an intervention is or is not effective without describing the type of participant on which the intervention was tested. The description requires specification of criteria for eligibility. This chapter focuses on how to define the study population. In addition, it considers two questions. First, to what extent will the results of the trial be generalizable to a broader population? Second, what impact does selection of eligibility criteria have on participant recruitment, or, more generally, study feasibility? This issue is also discussed in Chap. 10.

Fundamental Point

The study population should be defined in advance, stating unambiguous inclusion (eligibility) criteria. The impact that these criteria will have on study design, ability to generalize, and participant recruitment must be taken into account.

Definition of Study Population

The study population is the subset of the population with the condition or characteristics of interest defined by the eligibility criteria. The group of participants actually studied in the trial is selected from the study population. See Fig. 4.1.

Rationale

In reporting the study, the investigator needs to say what people were studied and how they were selected. The reasons for this are several. First, if an intervention is shown to be successful or unsuccessful, the medical and scientific communities must know to what kinds of people the findings apply.

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_4, © Springer Science+Business Media, LLC 2010

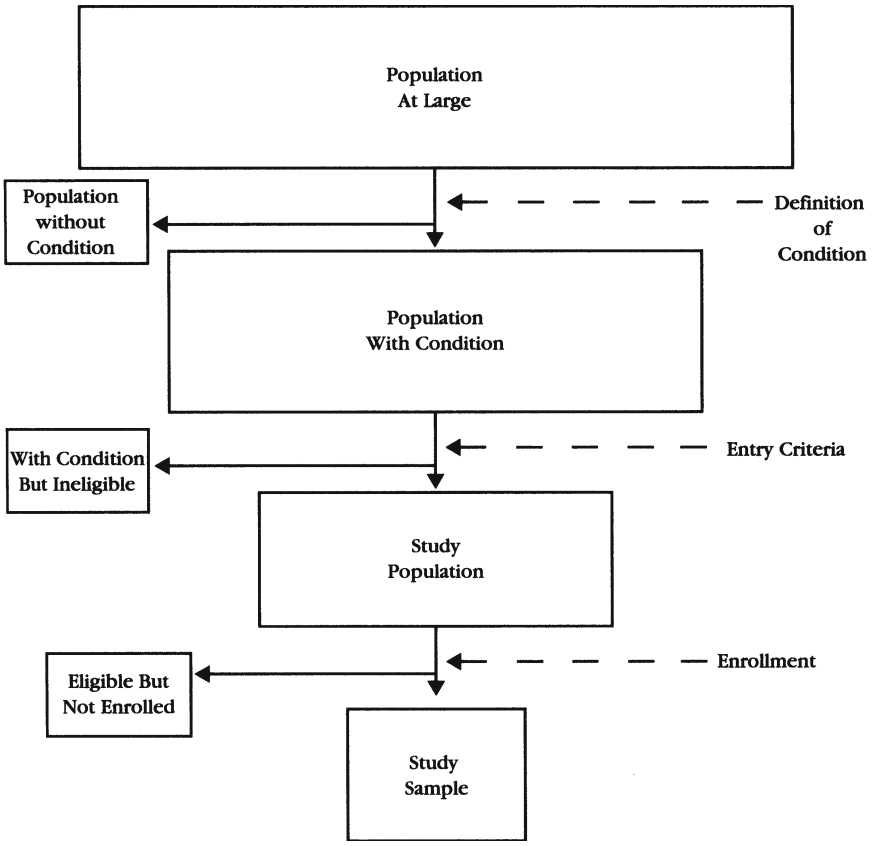


Fig. 4.1 Relationship of study sample to study population and general population (those with and those without the condition under study)

Second, knowledge of the study population helps other investigators assess the study's merit and appropriateness. For example, an antianginal drug may be found to be ineffective. Close examination of the description of the study population, however, could reveal that the participants represented a variety of ill-defined conditions characterized by chest pain. Thus, the study may not have been properly designed to evaluate the antianginal effects of the agent. Unfortunately, despite guidelines for reporting trial results [1], many publications contain inadequate characterization of the study participants [2]. Therefore, readers may be unable to assess fully the merit or applicability of the studies.

Third, for other investigators to be able to replicate the study, they need data descriptive of those enrolled. A similar issue is sometimes found in laboratory research. Because of incomplete discussion of details of the methods, procedures and preparation of materials, other investigators find it difficult to replicate an experiment. Before most research findings are widely accepted, they need to be

confirmed by independent scientists. Only small trials are likely to be repeated, but these are the ones, in general, that most need confirmation.

Considerations in Defining the Study Population

Inclusion criteria and reasons for their selection should be stated in advance. Ideally, all eligibility criteria should be precisely specified, but this is often impractical. Therefore, those criteria central to the study should be the most carefully defined. For example, in a study of survivors of a myocardial infarction, the investigator may be interested in excluding people with severe hypertension. He will require an explicit definition of myocardial infarction, but with regard to hypertension, it may be sufficient to state that people with a systolic or diastolic blood pressure above a specified level will be excluded. Note that even here, the definition of severe hypertension, though arbitrary, is fairly specific. In a study of antihypertensive agents, however, the above definition of severe hypertension is inadequate. If the investigator wants to include only people with diastolic blood pressure over 90 mmHg, he should specify how often it is to be determined, over how many visits, when, with what instrument, by whom, and in what circumstances. It may also be important to know which, if any, antihypertensive agents participants were on before entering the trial. For any study of antihypertensive agents, the criterion of hypertension is central; a detailed definition of myocardial infarction, on the other hand, may be less important.

If age is a restriction, the investigator should ideally specify not only that a participant must be over age 21, for example, but *when* he must be over 21. If a subject is 20 at the time of a prebaseline screening examination, but 21 at baseline, is he eligible? This should be clearly indicated. If diabetes is an exclusion criterion, is this only insulin-dependent diabetes or all diabetes? Does glucose intolerance warrant exclusion? How is diabetes defined? Often there are no "correct" ways of defining inclusion and exclusion criteria and arbitrary decisions must be made. Regardless, they need to be as clear as possible, and, when appropriate, with complete specifications of the technique and laboratory methods.

As discussed in Chap. 5, many clinical trials are of the "large, simple" model. In such trials, not only are the interventions relatively easy to implement, and the baseline and outcome variables limited, so too are the eligibility criteria. Definitions of eligibility criteria may not require repeated visits or special procedures. They may rely on previously measured variables that are part of a diagnostic evaluation, or on variables that are measured using any of several techniques, or on investigator judgment. For example, a detailed definition of myocardial infarction or hypertension may be replaced with, "Does the investigator believe a myocardial infarction has occurred?" or "Is hypertension present?" The advantage of this kind of criterion is its simplicity. The disadvantage is the possible difficulty that a clinician reading the results of the trial will have in deciding if the results are applicable to specific patients under her care. It should be noted, however, that even with the large simple trial model, the criteria are selected and specified in advance.

In general, eligibility criteria relate to participant safety and anticipated effect of the intervention. The following categories outline the framework upon which to develop individual criteria:

1. Participants who have the potential to benefit from the intervention are obviously candidates for enrollment into the study. The investigator selects participants on the basis of his scientific knowledge and the expectation that the intervention will work in a specific way on a certain kind of participant. For example, participants with a urinary infection are appropriate to enroll in a study of a new antibiotic agent known to be effective in vitro against the identified microorganism and thought to penetrate to the site of the infection in sufficient concentration. It should be evident from this example that selection of the participant depends on knowledge of the presumed mechanism of action of the intervention. Knowing at least something about the mechanism of action may enable the investigator to identify a well-defined group of participants likely to respond to the intervention. Thus, people with similar characteristics with respect to the relevant variable, that is, a *homogeneous* population, can be studied. In the above example, participants are homogeneous with regard to the type and strain of bacteria, and to site of infection. If age or renal or liver function is also critical, these too might be considered, creating an even more highly selected group.

   Even if the mechanism of action of the intervention is known, however, it may not be feasible to identify a homogeneous population because the technology to do so may not be available. For instance, the causes of headache are numerous and, with few exceptions, not easily or objectively determined. If a potential therapy were developed for one kind of headache, it would be difficult to identify precisely the people who might benefit.

   If the mechanism of action of the intervention is unclear, or if there is uncertainty at which stage of a disease a treatment might be most beneficial, a specific group of participants likely to respond cannot easily be selected. The Diabetic Retinopathy Study [3] evaluated the effects of photocoagulation on progression of retinopathy. In this trial, each person had one eye treated while the other eye served as the control. Participants were subgrouped on the basis of existence, location, and severity of vessel proliferation. Before the trial was scheduled to end, it became apparent that treatment was dramatically effective in the four most severe of the ten subgroups. To have initially selected for study only those four subgroups who benefited was not possible given existing knowledge.

   Some interventions may have more than one potentially beneficial mechanism of action. For example, if exercise reduces mortality or morbidity, is it because of its effect on cardiac performance, its weight-reducing effect, its effect on the person's sense of well-being, some combination of these effects, or some as yet unknown effect? The investigator could select study participants who have poor cardiac performance, or who are obese or who, in general, do not feel well. If he chooses incorrectly, his study would not yield a positive result. If he chose participants with all three characteristics and then showed benefit from exercise, he would never know which of the three aspects was important.

One could, of course, choose a study population, the members of which differ in one or more identifiable aspects of the condition being evaluated, i.e., a heterogeneous group. These differences could include stage or severity of a disease, etiology, or demographic factors. In the above exercise example, studying a heterogeneous population may be preferable. By comparing outcome with presence or absence of initial obesity or sense of well-being, the investigator may discover the relevant characteristics and gain insight into the mechanism of action. Also, when the study group is too restricted, there is no opportunity to discover whether an intervention is effective in a subgroup not initially considered. The broadness of the Diabetic Retinopathy Study was responsible for showing, after longer follow-up, that the remaining six subgroups also benefited from therapy [4]. If knowledge had been more advanced, only the four subgroups with the most dramatic improvement might have been studied. Obviously, after publication of the results of these four subgroups, another trial might have been initiated. However, valuable time would have been wasted. Extrapolation of conclusions to milder retinopathy might even have made a second study difficult. Of course, the effect of the intervention on a heterogeneous group may be diluted and the ability to detect a benefit may be reduced. That is the price to be paid for incomplete knowledge about mechanism of action.

Large, simple trials are, by nature, more heterogeneous in their study populations, than other sorts of trials. There is a greater chance that the participants will more closely resemble the mix of patients in many clinical practices. It is assumed, in the design, that the intervention affects a diverse group, and that despite such diversity, the effect of the intervention is more similar among the various kinds of participants than not.

Homogeneity and heterogeneity are matters of degree and knowledge. As scientific knowledge advances, ability to classify is improved. Today's homogeneous group may be considered heterogeneous tomorrow. The discovery of Legionnaires' disease [5], as a separate entity, caused by a specific organism improved possibilities for categorizing respiratory disease. Presumably, until that discovery, people with Legionnaires' disease were simply lumped together with people having other respiratory ailments.

2. In selecting participants to be studied, not only does the investigator require people in whom the intervention might work but he also wants to choose people in whom there is a high likelihood that he can detect the hypothesized results of the intervention. Careful choice will enable investigators to detect results in a reasonable period of time, given a reasonable number of participants and a finite amount of money.

For example, in a trial of an antianginal agent, an investigator would not wish to enroll a person who, in the past 2 years, has had only one brief angina pectoris episode (assuming such a person could be identified). The likelihood of finding an effect of the drug on this person is limited, since his likelihood of having many angina episodes during the expected duration of the trial is small. Persons with frequent episodes would be more appropriate. Similarly, many people

accept the hypothesis that, at least until it reaches a quite low level, LDL-cholesterol is a continuous variable in its impact on the risk of developing cardiovascular disease. Theoretically, an investigator could take almost any population with moderate or even relatively low LDL-cholesterol, attempt to lower it, and see if occurrence of cardiovascular disease is reduced. However, this would require studying an impossibly large number of people, since the calculation of sample size (Chap. 8) takes into account expected frequency of the primary response variables. When the expected frequency in the control group is low, as it would likely be in people who do not have elevated serum cholesterol, the number of people studied must be correspondingly high. From a sample size point of view it is, therefore, desirable to begin studying people with greater levels of risk factors and a consequent high expected event rate. If results from a first trial are positive, the investigator can then go to groups with lower levels. The initial Veterans Administration study of the treatment of hypertension [6] involved people with diastolic blood pressure from 115 to 129 mmHg. After therapy was shown to be beneficial in that group, a second trial was undertaken using people with diastolic blood pressures from 90 to 114 mmHg [7]. The latter study suggested that treatment should be instituted for people with diastolic blood pressure over 104 mmHg. Results were less clear for people with lower blood pressure. Subsequently, the Hypertension Detection and Follow-up Program [8] demonstrated benefit from treatment for people with diastolic blood pressure of 90 mmHg or above.

Sometimes, it may be feasible to enroll people with low levels of a risk factor if other characteristics elevate the absolute risk. For example, the Justification for the Use of Statins in Prevention: an Intervention Trial Evaluating Rosuvastatin (JUPITER) [9] used C-reactive protein to select those with LDL-cholesterol levels under 130 mg/dl (3.4 mmol/l) but who were likely to be at higher risk of developing coronary heart disease. The cholesterol-lowering agent rosuvastatin was shown to significantly lower the incidence of CHD.

Generally, if the primary response is continuous (e.g., blood pressure, blood sugar, body weight), change is easier to detect when the initial level is extreme. In a study to see whether a new drug is antihypertensive, there might be a more pronounced drop of blood pressure in a participant with diastolic pressure of 100 mmHg or greater than in one with diastolic pressure of 90 mmHg or less. There are exceptions to this rule, especially if a condition has multiple causes. The relative frequency of each cause might be different across the spectrum of values. For example, genetic disorders might be heavily represented among people with extremely high LDL-cholesterol. These lipid disorders may require alternative therapies or may even be resistant to usual methods of reducing LDL-cholesterol. In addition, use of participants with lower levels of a variable such as cholesterol might be less costly [10]. This is because of lower screening costs. Therefore, while in general, use of higher risk participants is preferable, other considerations can modify this.

3. Most interventions are likely to have adverse events. The investigator needs to weigh these against possible benefit when he evaluates the feasibility of doing the study. However, any person for whom the intervention is known to be harmful

should not, except in unusual circumstances, be admitted to the trial. Pregnant women are often excluded from drug trials (unless, of course, the primary question concerns pregnancy). The amount of additional data obtained may not justify the risk of possible teratogenicity. Similarly, investigators would probably exclude from a study of almost any of the anti-inflammatory drugs people with a recent history of gastric bleeding. Gastric bleeding is a fairly straightforward and absolute contraindication for enrollment. Yet, an exclusion criterion such as "history of major gastric bleed," leaves much to the judgment of the investigator. The word "major" implies that gastric hemorrhaging is not an absolute contraindication, but a relative one that depends upon clinical judgment. The phrase also recognizes the question of anticipated risk vs. benefit, because it does not clearly prohibit people with a mild bleeding episode in the distant past from being placed on an anti-inflammatory drug. It may very well be that such people take aspirin or similar agents – possibly for a good reason – and studying such people may prove more beneficial than hazardous.

Note that these exclusions apply only before enrollment into the trial. During a trial participants may develop symptoms or conditions which would have excluded them had any of these conditions been present earlier. In these circumstances, the participant may be removed from the intervention regimen if it is contraindicated, but she should be kept in the trial for purposes of analysis. As described in Chap. 17, being off the intervention does not mean that a participant is out of the trial.

4. The issue of competing risk is generally of greater interest in long-term studies. Participants at high risk of developing conditions, which preclude the ascertainment of the event of interest, should be excluded from enrollment. The intervention may or may not be efficacious in such participants, but the necessity for excluding them from enrollment relates to design considerations. In many studies of people with heart disease, those who have cancer or severe kidney or liver disorders are excluded because these diseases might cause the participant to die or withdraw from the study before the primary response variable can be observed. However, even in short-term studies, the competing risk issue needs to be considered. For example, an investigator may be studying a new intervention for a specific congenital heart defect in infants. Such infants are also likely to have other life-threatening defects. The investigator would not want to enroll infants if one of these other conditions were likely to lead to the death of the infant before he had an opportunity to evaluate the effect of the intervention. This matter is similar to the one raised in Chap. 3, which presented the problem of the impact of high expected total mortality on a study in which the primary response variable is morbidity or cause-specific mortality. When there is competing risk, the ability to assess the true impact of the intervention is, at best, lessened. At worst, if the intervention somehow has either a beneficial or harmful effect on the coexisting condition, biased results for the primary question can be obtained.
5. Investigators prefer, ordinarily, to enroll only participants who are likely to adhere to the study protocol. Participants are expected to take their assigned

intervention (usually a drug) and return for scheduled follow-up appointments regardless of the intervention assignment. In unblinded studies, participants are asked to accept the random assignment, even after knowing its identity, and abide by the protocol. Moreover, participants should not receive the study intervention from sources outside the trial during the course of the study. Participants should also refrain from using other interventions that may compete with the study intervention. Nonadherence by participants reduces the opportunity to observe the true effect of intervention. Unfortunately, there are no failsafe ways of selecting perfect participants. Traditional guidelines have led to disappointing results. For a further discussion of adherence, see Chap. 14.

An exception to this effort to exclude those less likely to take their medication or otherwise comply with the protocol is what some have termed "pragmatic" clinical trials [11]. These trials are meant to mimic real-world practice, with inclusion of participants who may fail to adhere consistently to the intervention. To compensate for the lower expected difference between the intervention and control groups, these trials need to be quite big, and have other characteristics of large, simple trials.

It should be noted that cultural or political issues, in addition to scientific, public health, or study design considerations, may affect selection of the study populations. Some have argued that too many clinical trials excluded, for example, women, the elderly, or minority groups, or that even if not excluded, insufficient attention was paid to enrolling them in adequate numbers [12–14]. Policies from the U.S. National Institutes of Health now require clinical trials to include certain groups in enough numbers to allow for "valid analysis" [15]. The effect of these kinds of policies on eligibility criteria, sample size, and analysis must be considered when designing a trial.

Generalization

Study samples or participants are usually nonrandomly chosen from the study population, which in turn is defined by the eligibility criteria (Fig. 4.1). As long as selection of participants into a trial occurs, and as long as enrollment is voluntary, participants must be regarded as special and not truly representative of the study population. Therefore, investigators have the problem of generalizing from participants actually in the trial to the study population and then to the population with the condition. Some have termed using trial results to draw conclusions about the broader population as "external validity." Defined medical conditions and quantifiable or discrete variables, such as age, sex, or elevated blood sugar, can be clearly stated and measured. For these characteristics, specifying in what way the study participants and study population are different from the population with the condition is relatively easy. Judgments about the appropriateness of generalizing study results can, therefore, be made. Other factors of the study participants are less easily characterized. Obviously, an investigator studies only those participants available to him.

If he lives in Florida, he will not be studying people living in Maine. Even within a geographical area, many investigators are based at hospitals or universities. Furthermore, many hospitals are referral centers. Only certain types of participants come to the attention of investigators at these institutions. It may be impossible to decide whether these factors are relevant when generalizing to other geographical areas or patient care settings. Multicenter trials typically enhance the ability to generalize. The growth of international trials, however, raises the issue of relevance of results from geographical areas with very different clinical care systems.

Many trials now involve participants from community or practice-based settings. Results from these "practical" or "pragmatic" trials may more readily be translated to the broader population. Even here, however, those who choose to become investigators likely differ from other practitioners in the kinds of patients they see.

It is often forgotten that participants must agree to enroll in a study. What sort of person volunteers for a study? Why do some agree to participate while others do not? The requirement that study participants sign informed consent or return for periodic examinations is sufficient to make certain people unwilling to participate. Sometimes the reasons are not obvious. What is known, however, is that volunteers can be different from non-volunteers [16–18]. They are usually in better health and are more likely to comply with the study protocol. However, the reverse could also be true. A person might be more motivated if she has disease symptoms. In the absence of knowing what motivates the particular study participants, appropriate compensatory adjustments cannot be made in the analysis. Because specifying how volunteers differ from others is difficult, an investigator cannot confidently identify those segments of the study population or the general population that these study participants supposedly represent. (See Chap. 10 for a discussion of factors that people cite for enrolling or not enrolling in trials.)

One approach to addressing the question of representativeness is to maintain a log or registry, which lists prospective participants identified, but not enrolled, and the reasons for excluding them. This log can provide an estimate of the proportion of all potentially eligible people who meet study entrance requirements and can also indicate how many otherwise eligible people refused enrollment. In an effort to further assess the issue of representativeness, response variables in those excluded have also been monitored. In the Norwegian Multicenter Study of timolol [19], people excluded because of contraindication to the study drug or competing risks had a mortality rate twice that of those who enrolled. The Coronary Artery Surgery Study included a randomized trial that compared coronary artery bypass surgery against medical therapy and a registry of people eligible for the trial but who refused to participate [20]. The enrolled and not enrolled groups were alike in most identifiable respects. Survival in the participants randomly assigned to medical care was the same as those receiving medical care but not in the trial. The findings for those undergoing surgery were similar. Therefore, in this particular case, the trial participants appeared to be representative of the study population.

With more attention being paid to privacy issues, however, it may not be possible to assess outcomes in those not agreeing to enter a trial. Some people may

consent to allow follow-up, even if they do not enroll, but many will not. Thus, comparison of trial results with results in those refusing to enter a trial, in an effort to show that the trial can be generalized, may prove difficult.

Since the investigator can describe only to a limited extent the kinds of participants in whom an intervention was evaluated, a leap of faith is always required when applying any study findings to the population with the condition. In taking this jump, one must always strike a balance between making unjustifiably broad generalizations and being too conservative in one's claims. Some extrapolations are reasonable and justifiable from a clinical point of view, especially in the light of subsequent information.

Many trials of aspirin and other anti-platelet agents in those who have had a heart attack have shown that these agents reduce recurrent myocardial infarction and death in both men and women [21]. The Physicians' Health Study, conducted in the 1980s, concluded that aspirin reduced myocardial infarction in men over age 50 without previously documented heart disease [22]. Although it was reasonable to expect that a similar reduction would occur in women, it was unproven. Importantly, aspirin was shown in the Physicians' Health Study and elsewhere [23] to increase hemorrhagic stroke. Given the lower risk of heart disease in premenopausal women, whether the trade-off between adverse effects and benefit was favorable was far from certain. The U.S. Food and Drug Administration approved aspirin for primary prevention in men, but not women. The Women's Health Study was conducted in the 1990s and early 2000s [24]. Using a lower dose of aspirin than was used in the Physicians' Health Study, it found evidence of benefit on heart disease only in women at least 65 years old. Based on that, generalization of the Physicians' Health Study results to primary prevention in all women would not have been prudent. A subsequent meta-analysis, however, suggested that the benefits of aspirin for primary prevention were similar in women and men. We must always be open to consider new information in our interpretation of study results [25].

Recruitment

The impact of eligibility criteria on recruitment of participants should be considered when deciding on these criteria. Using excessive restrictions in an effort to obtain a pure (or homogeneous) sample can lead to extreme difficulty in obtaining sufficient participants and may raise questions regarding generalization of the trial results. Age and sex are two criteria that have obvious bearing on the ease of enrolling subjects. The Coronary Primary Prevention Trial undertaken by the Lipid Research Clinics was a collaborative trial evaluating a lipid-lowering drug in men between the ages of 35 and 59 with severe hypercholesterolemia. One of the Lipid Research Clinics [26] noted that approximately 35,000 people were screened and only 257 participants enrolled. Exclusion criteria, all of which were perfectly reasonable and scientifically sound, coupled with the number of people who refused to enter the study, brought the overall trial yield down to less than 1%. As

discussed in Chap. 10, this example of greater than expected numbers being screened, as well as unanticipated problems in reaching potential participants, is common to most clinical trials.

If entrance criteria are properly determined in the beginning of a study, there should be no need to change them unless interim results suggest harm in a specific subgroup (see Chap. 16). As discussed earlier in this chapter, eligibility criteria are appropriate if they exclude those who might be harmed by the intervention, those who are not likely to be benefited by the intervention, or those who are not likely to comply with the study protocol. The reasons for each criterion should be carefully examined during the planning phase of the study. If they do not fall into one of the above categories, they should be reassessed. Whenever an investigator considers changing criteria, he needs to look at the effect of changes on participant safety and study design. It may be that, in opening the gates to accommodate more participants, he increases the required sample size, because the participants admitted may have lower probability of developing the primary response variable. He can thus lose the benefits of added recruitment. In summary, capacity to recruit participants and to carry out the trial effectively could greatly depend on the eligibility criteria that are set. As a consequence, careful thought should go into establishing them.

References

1. CONSORT. http://www.consort-statement.org
2. Van Spall HGC, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. *JAMA* 2007;297:1233–1240.
3. Diabetic Retinopathy Study Research Group. Preliminary report on effects of photocoagulation therapy. *Am J Ophthalmol* 1976;81:383–396.
4. Diabetic Retinopathy Study Research Group. Photocoagulation treatment of proliferative diabetic retinopathy: the second report of diabetic retinopathy study findings. *Ophthalmology* 1978;85:82–106.
5. Fraser DW, Tsai TR, Orenstein W, et al. Legionnaires' Disease: description of an epidemic of pneumonia. *N Engl J Med* 1977;297:1189–1197.
6. Veterans Administration Cooperative Study Group on Antihypertensive Agents. Effects of treatment on morbidity in hypertension: results in patients with diastolic blood pressures averaging 115 through 129 mm Hg. *JAMA* 1967;202:1028–1034.
7. Veterans Administration Cooperative Study Group on Antihypertensive Agents. Effects of treatment on morbidity in hypertension: II. Results in patients with diastolic blood pressure averaging 90 through 114 mm Hg. *JAMA* 1970;213:1143–1152.
8. Hypertension Detection and Follow-up Program Cooperative Group. Five-year findings of the hypertension detection and follow-up program. 1. Reduction in mortality of persons with high blood pressure, including mild hypertension. *JAMA* 1979;242:2562–2571.
9. Ridker PM, Danielson E, Fonseca FAH, et al. Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *N Engl J Med* 2008;359:2195–2207.
10. Sondik EJ, Brown BW, Jr., Silvers A. High risk subjects and the cost of large field trials. *J Chronic Dis* 1974;27:177–187.

11. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA* 2003;290:1624–1632.
12. Douglas PS. Gender, cardiology, and optimal medical care. *Circulation* 1986;74:917–919.
13. Bennett JC, for the Board on Health Sciences Policy of the Institute of Medicine. Inclusion of women in clinical trials – policies for population subgroups. *N Engl J Med* 1993;329:288–292.
14. Freedman LS, Simon R, Foulkes MA, et al. Inclusion of women and minorities in clinical trials and the NIH Revitalization Act of 1993 – the perspective of NIH clinical trialists. *Control Clin Trials* 1995;16:277–285.
15. NIH Policy and Guidelines on the Inclusion of Women and Minorities as Subjects in Clinical Research – Amended, October, 2001. http://grants.nih.gov/grants/funding/women_min/guidelines_amended_10_2001.htm
16. Horwitz O, Wilbek E. Effect of tuberculosis infection on mortality risk. *Am Rev Respir Dis* 1971;104:643–655.
17. Wilhelmsen L, Ljungberg S, Wedel H, Werko L. A comparison between participants and non-participants in a primary preventive trial. *J Chronic Dis* 1976;29:331–339.
18. Smith P, Arnesen H. Mortality in non-consenters in a post-myocardial infarction trial. *J Intern Med* 1990;228:253–256.
19. Pedersen TR. The Norwegian Multicenter Study of timolol after myocardial infarction. *Circulation* 1983;67(suppl 1):I-49–I-53.
20. CASS Principal Investigators and Their Associates. Coronary Artery Surgery Study (CASS): a randomized trial of coronary artery bypass surgery. Comparability of entry characteristics and survival in randomized patients and nonrandomized patients meeting randomization criteria. *J Am Coll Cardiol* 1984;3:114–128.
21. Antithrombotic Trialists' Collaboration. Collaborative meta-analysis of randomised clinical trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients. *BMJ* 2002;324:71–86; correction *BMJ* 2002;324:141.
22. Steering Committee of the Physicians' Health Study Research Group. Final report on the aspirin component of the ongoing Physicians' Health Study. *N Engl J Med* 1989;321:129–135.
23. Peto R, Gray R, Collins R, et al. Randomized trial of prophylactic daily aspirin in British male doctors. *Br Med J* 1988;296:313–316.
24. Ridker PM, Cook NR, Lee I-M, et al. A randomized trial of low-dose aspirin in the primary prevention of cardiovascular disease in women. *N Engl J Med* 2005;352:1293–1304.
25. Berger JS, Roncaglioni MC, Avanzini F, et al. Aspirin for the primary prevention of cardiovascular events in women and men: a sex-specific meta-analysis of randomized controlled trials. *JAMA* 2006;295:306–313; correction *JAMA* 2006;295:2002.
26. Benedict GW. LRC Coronary Prevention Trial: Baltimore. *Clin Pharmacol Ther* 1979; 25:685–687.

Chapter 5
Basic Study Design

The foundations for the design of controlled experiments were established for agricultural application. They are described in several classical statistics textbooks [1–4]. From these sources evolved the basic design of controlled clinical trials.

Although the history of clinical experimentation contains several instances in which the need for control groups has been recognized [5, 6], this need was not widely accepted until the 1950s [7]. In the past, when a new intervention was first investigated, it was likely to be given to only a small number of people, and the outcome compared, if at all, to that in people with the same condition previously treated in a different manner. The comparison was informal and frequently based on memory alone. Sometimes, in one kind of what is sometimes called a "quasi-experimental" study, people were evaluated initially and then reexamined after an intervention had been introduced. In such studies, the changes from the initial state were used as the measure of success or failure of the new intervention. What could not be known was whether the person would have responded in the same manner if there had been no intervention at all. However, then – and sometimes even today – this kind of observation has formed the basis for the use of new interventions.

Of course, some results are so highly dramatic that no comparison group is needed. Successful results of this magnitude, however, are rare. One example is the effectiveness of penicillin in pneumococcal pneumonia. Another example originated with Pasteur who, in 1884, was able to demonstrate that a series of vaccine injections protected dogs from rabies [8]. He suggested that due to the long incubation time, prompt vaccination of a human being after infection might prevent the fatal disease. The first patient was a 9-year-old boy who had been bitten 3 days earlier by a rabid dog. The treatment was completely effective. Confirmation came from another boy who was treated within 6 days of having been bitten. During the next few years, hundreds of patients were given the anti-rabies vaccine. If given within certain time-limits, it was almost always effective.

Gocke reported on a similar, uncontrolled study of patients with acute fulminant viral hepatitis [9]. Nine consecutive cases had been observed, all of whom had a fatal outcome. The next diagnosed case, a young staff nurse in hepatic coma, was given immunotherapy in addition to standard treatment. The patient survived as did four others among eight given the antiserum. The author initially thought that this

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_5, © Springer Science+Business Media, LLC 2010

uncontrolled study was conclusive. However, in considering other explanations for the encouraging findings, he could not eliminate the possibility that a tendency to treat patients earlier in the course and more intensive care might be responsible for the observed outcome. Thus, he joined a double-blind, randomized trial comparing hyperimmune anti-Australia globulin to normal human serum globulin in patients with severe acute hepatitis. Nineteen of 28 patients (67.9%) randomized to control treatment died, compared to 16 of 25 patients (64%) randomized to treatment with exogenous antibody, a statistically nonsignificant difference [10].

A number of medical conditions are either of short duration or episodic in nature. Evaluation of therapy in these cases can be difficult in the absence of controlled studies. Snow and Kimmelman reviewed various uncontrolled studies of surgical procedures for Meniere's disease [11]. They found that about 75% of patients improved, but noted that this is similar to the 70% remission rate occurring without treatment.

Given the wide spectrum of the natural history of almost any disease and the variability of an individual's response to an intervention, most investigators recognize the need for a defined control or comparison group.

Fundamental Point

Sound scientific clinical investigation almost always demands that a control group be used against which the new intervention can be compared. Randomization is the preferred way of assigning participants to control and intervention groups.

Statistics and epidemiology textbooks and papers [12–31] cover various study designs in some detail. Green and Byar also present a "hierarchy of strength of evidence concerning efficacy of treatment" [32]. In their scheme, anecdotal case reports are weakest and confirmed randomized clinical trials are strongest, with various observational and retrospective designs in between. This chapter discusses several major clinical trial designs.

Most trials use the so-called parallel design. That is, the intervention and control groups are followed simultaneously from the time of allocation to one or the other. Exceptions to the simultaneous follow-up are historical control studies. These compare a group of participants on a new intervention with a previous group of participants on standard or control therapy. A modification of the parallel design is the cross-over trial, which uses each participant at least twice, at least once as a member of the control group and at least once as a member of one or more intervention groups. Another modification is a withdrawal study, which starts with all participants on the active intervention and then, usually randomly, assigns a portion to be followed on the active intervention and the remainder to be followed off the intervention. Factorial design trials, as described later in this chapter, employ two or more independent assignments to intervention or control.

Regardless of whether the trial is a typical parallel design or some variant, one must select the kind of control group and the way participants are allocated to

intervention or control. Controls may be on placebo, no treatment, usual or standard care, or a specified treatment. Randomized control and nonrandomized concurrent control studies both assign participants to either the intervention or the control group, but only the former makes the assignment by using a random procedure. Hybrid designs may use a combination of randomized and nonrandomized controls. Large, simple trials or pragmatic trials generally have broader and simpler eligibility criteria than other kinds of trials, but as with other studies, can use any of the indicated controls. Allocation to intervention or control may also be done differently, even if randomized. Randomization may be by individual participant or by groups of participants (group or cluster assignment). Adaptive designs may adjust intervention or control assignment or sample size on the basis of participant characteristics or outcomes.

Finally, there are superiority trials and equivalence or noninferiority trials. A superiority trial, which for many years was the typical kind of trial, assesses whether the new intervention is better or worse than the control. An equivalence trial would assess if the new intervention is more or less equal to the control. A noninferiority trial evaluates whether the new intervention is no worse than the control by some margin, delta (δ). In both of these latter cases, the control group would be on a treatment that had previously been shown to be effective, i.e., have an active control.

Questions have been raised concerning the method of selection of the control group, but the major controversy in the past revolved around the use of historical vs. randomized control [33–35]. With regard to drug evaluation, this controversy is less intense than in the past. It is still being hotly contested, however, in the evaluation of new devices or procedures [36, 37]. No study design is perfect or can answer all questions. Each of the designs has advantages and disadvantages, but a randomized control design is the standard by which other studies should be judged. A discussion of sequential designs is postponed until Chap. 16 because the basic feature involves interim analyses.

For each of the designs, it is assumed, for simplicity of discussion, that a single control group and a single intervention group are being considered. These designs can be extended to more than one intervention group and more than one control group.

Randomized Control Trials

Randomized control trials are comparative studies with an intervention group and a control group; the assignment of the subject to a group is determined by the formal procedure of randomization. Randomization, in the simplest case, is a process by which all participants are equally likely to be assigned to either the intervention group or the control group. The features of this technique are discussed in Chap. 6. There are three advantages of the randomized design over other methods for selecting controls [35].

First, randomization removes the potential of bias in the allocation of participants to the intervention group or to the control group. Such allocation bias could easily

occur, and cannot be necessarily prevented, in the nonrandomized concurrent or historical control study because the investigator or the participant may influence the choice of intervention. This influence can be conscious or subconscious and can be due to numerous factors, including the prognosis of the participant. The direction of the allocation bias may go either way and can easily invalidate the comparison.

The second advantage, somewhat related to the first, is that randomization tends to produce comparable groups; that is, measured as well as unknown or unmeasured prognostic factors and other characteristics of the participants at the time of randomization will be, on the average, evenly balanced between the intervention and control groups. This does not mean that in any single experiment all such characteristics, sometimes called baseline variables or covariates, will be perfectly balanced between the two groups. However, it does mean that for independent covariates, whatever the detected or undetected differences that exist between the groups, the overall magnitude and direction of the differences will tend to be equally divided between the two groups. Of course, many covariates are strongly associated; thus, any imbalance in one would tend to produce imbalances in the others. As discussed in Chaps. 6 and 17, stratified randomization and stratified analysis are methods commonly used to guard against and adjust for imbalanced randomizations.

The third advantage of randomization is that the validity of statistical tests of significance is guaranteed. As has been stated [35], "although groups compared are never perfectly balanced for important covariates in any single experiment, the process of randomization makes it possible to ascribe a probability distribution to the difference in outcome between treatment groups receiving equally effective treatments and thus to assign significance levels to observed differences." The validity of the statistical tests of significance is not dependent on the balance of the prognostic factors between the two groups. The chi-square test for two-by-two tables and Student's *t*-test for comparing two means can be justified on the basis of randomization alone without making further assumptions concerning the distribution of baseline variables. If randomization is not used, further assumptions concerning the comparability of the groups and the appropriateness of the statistical models must be made before the comparisons will be valid. Establishing the validity of these assumptions may be difficult.

Randomized and nonrandomized trials of the use of anticoagulant therapy in patients with acute myocardial infarctions were reviewed by Chalmers et al. and the conclusions compared [38]. Of 32 studies, 18 used historical controls and involved a total of 900 patients, eight used nonrandomized concurrent controls and involved over 3,000 patients, and six were randomized trials with a total of over 3,800 patients. The authors reported that 15 of the 18 historical control trials and five of the eight nonrandomized concurrent control trials showed statistically significant results favoring the anticoagulation therapy. Only one of the six randomized control trials showed significant results in support of this therapy. Pooling the results of these six randomized trials yielded a statistically significant 20% reduction in total mortality, confirming the findings of the nonrandomized studies. Pooling the results of the nonrandomized control studies showed a reduction of about 50% in total mortality in the intervention groups, more than twice the decrease seen in the randomized trials. Peto [39] has

assumed that this difference in reduction is due to bias. He suggests that since the presumed bias in the nonrandomized trials was of the same order of magnitude as the presumed true effect, the nonrandomized trials could have yielded positive answers even if the therapy had been of no benefit. Of course, pooling results of several studies can be hazardous. As pointed out by Goldman and Feinstein [40], not all randomized trials of anticoagulants study the same kind of participants, use precisely the same intervention or measure the same response variables. And, of course, not all randomized trials are done equally well. The principles of pooled analysis, or meta-analysis, are covered in Chap. 17.

In the 1960s, Grace, Muench and Chalmers [41] reviewed studies involving portacaval shunt operations for patients with portal hypertension from cirrhosis. In their review, 34 of 47 nonrandomized studies strongly supported the shunt procedure, while only one of the four randomized control trials indicated support for the operation. The authors concluded that the operation should not be endorsed.

Sacks and coworkers expanded the work by Chalmers et al. referenced above, to five other interventions [42]. They concluded that selection biases led historical control studies to favor inappropriately the new interventions. It was also noted that many randomized control trials were of inadequate size, and therefore may have failed to find benefits that truly existed [43]. Chalmers and his colleagues also examined 145 reports of studies of treatment after myocardial infarction [44]. Of the 57 studies that used a randomization process with concealment of the intervention allocation, 14% had at least one significant ($p < 0.05$) maldistribution of baseline variables with 3.4% of all of the variables significantly different between treatment groups. Of these 57 studies, 9% found significant outcome differences between groups. This contrasted with 58% having baseline variable differences among the 43 reports where the control groups were selected by means of a nonrandom process, with 34% of all of the variables being significantly different between groups. The outcomes between groups were significantly different 58% of the time. For the 45 studies that used a randomized, but unblinded process to select the control groups, the results were in between; 28% had baseline imbalances, 7% of the baseline variables were significantly different, and 24% showed significant outcome differences. See Chap. 9 regarding testing for baseline differences.

The most frequent objections to the use of the randomized control clinical trial were stated by Ingelfinger [45], to be "emotional and ethical." Many clinicians feel that they must not deprive a participant from receiving a new therapy or intervention which they, or someone else, believe to be beneficial, regardless of the validity of the evidence for that claim. The argument aimed at randomization is that in the typical trial it deprives about one-half the participants from receiving the new and presumed better intervention. There is a large literature on the ethical aspects of randomization. See Chap. 2 for a discussion of this issue.

Not all clinical studies can use randomized controls. Occasionally, the prevalence of the disease is so rare that a large enough population cannot be readily obtained. In such an instance, only case–control studies might be possible. Such studies, which are not clinical trials according to the definition in this book, are discussed in standard epidemiology textbooks [15, 16, 22, 28].

Zelen proposed a modification of the standard randomized control study [46]. He argued that investigators are often reluctant to recruit prospective trial participants not knowing to which group the participant will be assigned. Expressing ignorance of optimal therapy compromises the traditional doctor–patient relationship. Zelen, therefore, suggested randomizing eligible participants before informing them about the trial. Only those assigned to active intervention would be asked if they wish to participate. The control participants would simply be followed and their outcome monitored. Obviously, such a design could not be blinded. Another major criticism of this controversial design centers around the ethical concern of not informing participants that they are enrolled in a trial. The efficiency of the design has also been evaluated [47]. It depends on the proportion of participants consenting to comply with the assigned intervention. To compensate for this possible inefficiency, one needs to increase the sample size (Chap. 8). The Zelen approach has been tried with varying degrees of success [48, 49]. Despite having been proposed in 1979, it does not appear to have been widely used.

Nonrandomized Concurrent Control Studies

Controls in this type of study are participants treated without the new intervention at approximately the same time as the intervention group is treated. Participants are allocated to one of the two groups, but by definition this is not a random process. An example of a nonrandomized concurrent control study would be a comparison of survival results of patients treated at two institutions, one institution using a new surgical procedure and the other using more traditional medical care.

To some investigators, the nonrandomized concurrent control design has advantages over the randomized control design. Those who object to the idea of ceding to chance the responsibility for selecting a person's treatment may favor this design. It is also difficult for some investigators to convince potential participants of the need for randomization. They find it easier to select a group of people to receive the intervention and would prefer to select the control group by means of matching key characteristics.

The major weakness of the nonrandomized concurrent control study is the potential that the intervention group and control group are not strictly comparable. It is difficult to prove comparability because the investigator must assume that she has information on all the important prognostic factors. Selecting a control group by matching on more than a few factors is impractical, and the comparability of a variety of other characteristics would still need to be evaluated. In small studies, an investigator is unlikely to find real differences, which may exist between groups before the initiation of intervention since there is poor sensitivity to detect such differences. Even for large studies that could detect most differences of real clinical importance, the uncertainty about the unknown or unmeasured factors is still of concern.

Is there, for example, some unknown and unmeasurable process that results in one type of participant's being recruited more often into one group and not into the other? If all participants come from one institution, physicians may select

participants into one group based on subtle and intangible factors. In addition, there exists the possibility for subconscious bias in the allocation of participants to either the intervention or control group. One group might come from a different socioeconomic class than the other group. All of these uncertainties will decrease the credibility of the concurrent but nonrandomized control study. For any particular question, the advantages of reduced cost, relative simplicity, and investigator and participant acceptance must be carefully weighed against the potential biases before a decision is made to use a nonrandomized concurrent control study. We believe this will occur very rarely.

Historical Controls and Databases

In historical control studies, a new intervention is used in a series of participants, and the results are compared to the outcome in a previous series of comparable participants. Historical controls are thus, by this definition, nonrandomized and nonconcurrent.

Strengths of Historical Control Studies

The argument for using a historical control design is that all new participants can receive the new intervention. As argued by Gehan and Freireich [33] many clinicians believe that no participant should be deprived of the possibility of receiving a new therapy or intervention. Some require less supportive evidence than others to accept a new intervention as being beneficial. If an investigator is already of the opinion that the new intervention is beneficial, then she would most likely consider any restriction on its use unethical. Therefore, she would favor a historical control study. In addition, participants may be more willing to enroll in a study if they can be assured of receiving a particular therapy or intervention. Finally, since all new participants will be on the new intervention, the time required to complete recruitment of participants for the trial will be cut approximately in half. This allows investigators to obtain results faster or do more studies with given resources. Alternatively, the sample size for the intervention group can be larger, with increased power.

Gehan emphasized the ethical advantages of historical control studies and pointed out that they have contributed to medical knowledge [50]. Lasagna argued that medical practitioners traditionally relied on historical controls when making therapeutic judgments. He maintained that, while sometimes faulty, these judgments are often correct and useful [51].

Typically, historical control data can be obtained from two sources. First, control group data may be available in the literature. These data are often undesirable because it is difficult, and perhaps impossible, to establish whether the control and intervention groups are comparable in key characteristics at the onset. Even if such

characteristics were measured in the same way, the information may not be published and for all practical purposes it will be lost. Second, data may not have been published but may be available on computer files or in medical charts. Such data on control participants, for example, might be found in a large center, which has several ongoing clinical investigations. When one study is finished, the participants in that study may be used as a control group for some future study. Centers that do successive studies, as in cancer research, will usually have a system for storing and retrieving the data from past studies for use at some future time. The advent of electronic medical records may also facilitate access to data from multiple sources although it does not solve the problem of nonstandard and variable assessment or missing information.

Limitations of Historical Control Studies

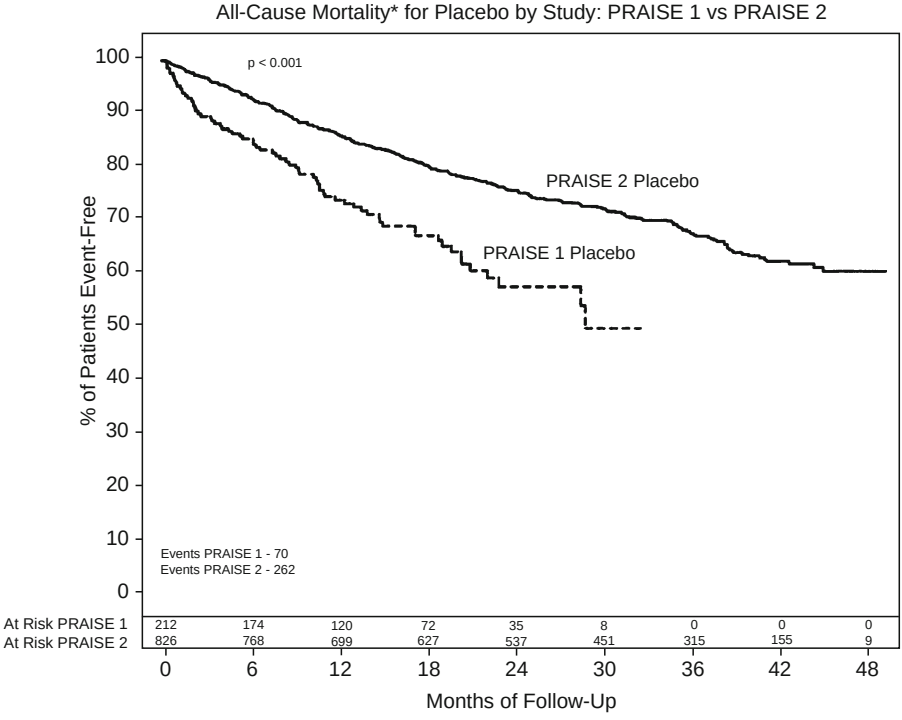
Despite the time and cost benefits, as well as the ethical considerations, historical control studies have potential limitations, which should be kept in mind. They are particularly vulnerable to bias. Moertel [52] cited a number of examples of treatments for cancer, which have been claimed, on the basis of historical control studies, to be beneficial. Many treatments in the past were declared breakthroughs on the basis of control data as old as 30 years. Pocock [53] identified 19 instances of the same intervention having been used in two consecutive trials employing similar participants at the same institution. Theoretically, the mortality in the two groups using the same treatment should be similar. Pocock noted that the difference in mortality rates between such groups ranged from –46% to +24%. Four of the 19 comparisons of the same intervention showed differences significant at the 5% level.

An improvement in outcome for a given disease may be attributed to a new intervention when, in fact, the improvement may stem from a change in the patient population or patient management. Shifts in patient population can be subtle and perhaps undetectable. In a Veterans Administration Urological Research Group study of prostate cancer [54], 2,313 people were randomized to placebo or estrogen treatment groups over a 7-year period. For those enrolled during the last 2–3 years, no differences were found between the placebo and estrogen groups. However, those assigned to placebo entering in the first 2–3 years had a shorter survival time than those assigned to estrogen entering in the last 2–3 years of the study. The reason for the early apparent difference is probably that the people randomized earlier were older than the later group and thus were at higher risk of death during the period of observation [35]. The results would have been misleading had this been a historical control study and had a concurrent randomized comparison group not been available.

A more recent example involves two trials evaluating the potential benefit of amlodipine, a calcium channel blocker, in patients with heart failure. The first trial, the Prospective Randomized Amlodipine Survival Evaluation, referred to as PRAISE-I [55], randomized participants to amlodipine or placebo, stratifying by

ischemic or nonischemic etiology of the heart failure. The primary outcome, death plus hospitalization for cardiovascular reasons, was not significantly different between groups ($p=0.31$), but the reduction in mortality almost reached significance ($p=0.07$). An interaction with etiology was noted, with all of the benefit from amlodipine in both the primary outcome and mortality seen in those with nonischemic etiology. A second trial, PRAISE-2 [56], was conducted in only those with nonischemic causes of heart failure. The impressive subgroup finding noted in PRAISE-1 were not replicated. Of relevance here is that the event rates in the placebo group in PRAISE-2 were significantly lower than in the nonischemic placebo participants from the first trial (see Fig. 5.1).

Even though the same investigators conducted both trials using the same protocol, the kinds of people who were enrolled into the second trial were markedly different from the first trial. Covariate analyses were unable to account for the difference in outcome.



Information for PRAISE 2 is from the ENDPT dataset sent to SDAC on December 19, 1999. The PRAISE 1 results are for the non-ischemic subgroup only.*For PRAISE 1, transplants have been censored at the time of transplant and are not considered an event for this analysis. For PRAISE 2, patients with transplants are followed for survival post-transplant.

Fig. 5.1 PRAISE 1 and 2 placebo arms

On a broader scale, for both known and unknown reasons, in many countries trends in prevalence of various diseases occur [57]. Therefore, any clinical trial in those conditions, involving long-term therapy using historical controls, would need to separate the treatment effect from the time trends, an almost impossible task. Examples are seen in Figs. 5.2 and 5.3.

Figure 5.2 illustrates the changes over time, in rates of the leading causes of death in the U.S. [58]. A few of the causes exhibit quite large changes. Figure 5.3

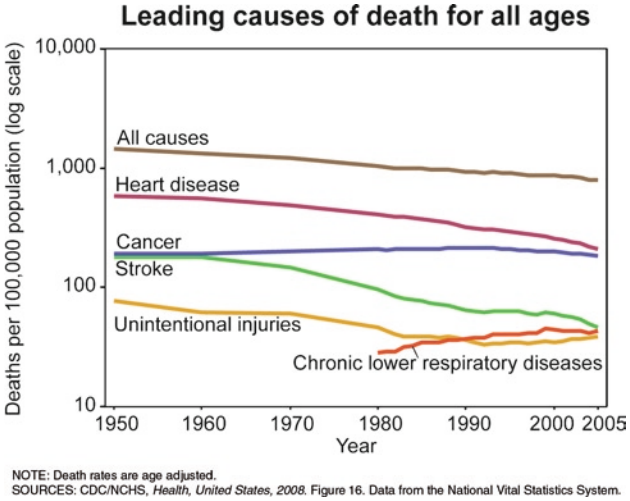


Fig. 5.2 Trends in causes of death in the U.S.

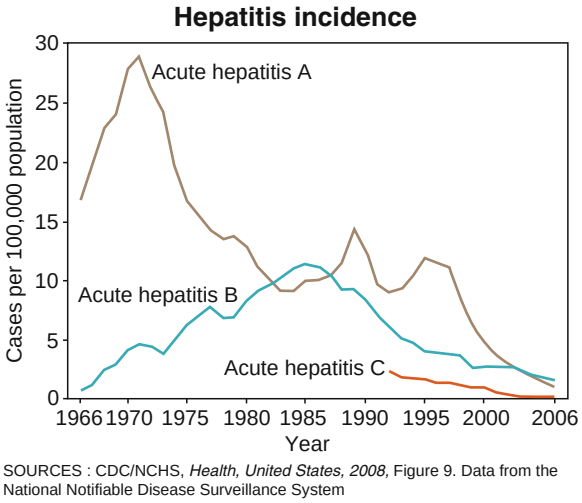


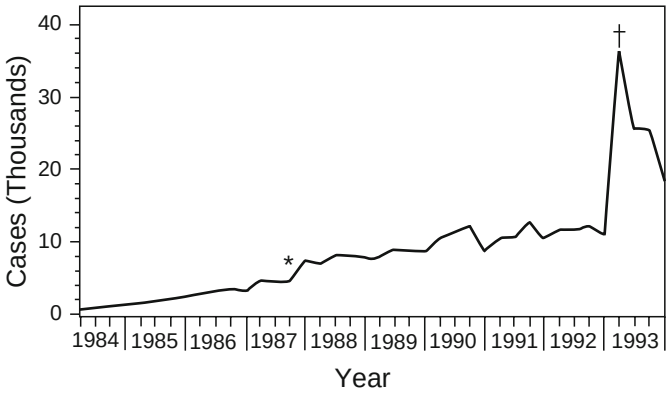
Fig. 5.3 Changes in incidence of hepatitis, by type, in the U.S.

shows incidence of hepatitis in the U.S. [58]. The big changes make interpretation of historical control trials difficult.

The method by which participants are selected for a particular study can have a large impact on their comparability with earlier participant groups or general population statistics. In the Coronary Drug Project [59], a trial of survivors of myocardial infarction initiated in the 1960s, an annual total mortality rate of 6% was anticipated in the control group based on rates from a fairly unselected group of myocardial infarction patients. In fact, a control group mortality rate of about 4% was observed, and no significant differences were seen between the intervention groups and the control group. Using the historical control approach, a 33% reduction in mortality might have been claimed for the treatments. One explanation for the discrepancy between anticipated and observed mortality is that entry criteria excluded those most seriously ill.

Shifts in diagnostic criteria for a given disease due to improved technology can cause major changes in the recorded frequency of the disease and in the perceived prognosis of the subjects with the disease. The use of elevated serum troponin, sometimes to the exclusion of the need for other features of a myocardial infarction such as symptoms or electrocardiographic changes, has clearly led to the ability to diagnose more infarctions. Conversely, the ability to abort an evolving infarction by means of percutaneous coronary intervention or thrombolytic therapy can reduce the number of clearly diagnosed infarctions.

In 1993, the Centers for Disease Control and Prevention (CDC) in the USA implemented a revised classification system for HIV infection and an expanded surveillance case definition of AIDS. This affected the number of cases reported [60, 61]. See Fig. 5.4.



Case definition revised in October 1987 to include additional illnesses and to revise diagnostic criteria (3).
† Case definition revised in 1993 to include CD4+ criteria and three illnesses (pulmonary tuberculosis, recurrent pneumonia, and invasive cervical cancer) (1).

Fig. 5.4 AIDS cases, by quarter year of report – United States, 1984–1993 [61]

International coding systems and names of diseases change periodically and, unless one is aware of the modifications, prevalence of certain conditions can appear to change abruptly. For example, when the Eighth Revision of the International Classification of Diseases came out in 1968, almost 15% more deaths were assigned to ischemic heart disease than had been assigned in the Seventh Revision [62]. When the Ninth Revision appeared in 1979, there was a correction downward of a similar magnitude [63].

A common concern about historical control designs is the accuracy and completeness with which control group data are collected. With the possible exception of special centers, which have many ongoing studies, data are generally collected in a nonuniform manner by numerous people with diverse interests in the information. Lack of uniform collection methods can easily lead to incomplete and erroneous records. Data on some important prognostic factors may not have been collected at all. Because of the limitations of data collected historically from medical charts, records from a center that conducts several studies and has a computerized data management system, may provide the most reliable historical control data.

Role of Historical Controls

Despite the limitations of the historical control study, it does have a place in scientific investigation. As a rapid, relatively inexpensive method of obtaining initial impressions regarding a new therapy, such studies can be important. This is particularly so if investigators understand the potential biases and are willing to miss effective new therapies if bias works in the wrong direction. Bailar et al. [64] identified several features, which can strengthen the conclusions to be drawn from historical control studies. These include an a priori identification of a reasonable hypothesis and advance planning for analysis.

In some special cases where the diagnosis of a disease is clearly established and the prognosis is well known or the disease highly fatal, a historical control study may be the only reasonable design. The results of penicillin in treatment of pneumococcal pneumonia were so dramatic in contrast to previous experience that no further evidence was really required. Similarly, the benefits of treatment of malignant hypertension became readily apparent from comparisons with previous, untreated populations [65–67].

The use of prospective registries to characterize patients and evaluate effects of therapy has been advocated [68–70]. Supporters say that a systematic approach to data collection and follow-up can provide information about the local patient population, and can aid in clinical decision making. They argue that clinical trial populations may not be representative of the patients actually seen by a physician. Moon et al. described the use of databases derived from clinical trials to evaluate therapy [71]. They stress that the high quality data obtained through these sources can reduce the problems of the typical historical control study. The use of databases has expanded

in recent years. Outcomes research has burgeoned because of the relative ease of accessing huge computerized medical databases [72]. The primary reasons have been the speed and lesser cost of such analyses, compared with clinical trials. Databases can also be used to identify adverse events. Examples are comparisons of different antihypertensive agents and risk of stroke [73] and cyclooxygenase 2 (COX 2) inhibitors and risk of coronary heart disease [74]. In addition, databases likely represent a much broader population than the typical clinical trial and can therefore complement clinical trial findings. This information can be useful as long as it is kept in mind that users and nonusers of a medication likely have different characteristics.

Others [32, 75–77] have emphasized limitations of registry studies such as potential bias in treatment assignment, multiple comparisons, lack of standardization in collecting and reporting data, and missing data. Another weakness of prospective database registries is that they rely heavily on the validity of the model employed to analyze the data [78].

There is no doubt that analyses of large databases can provide important information about disease occurrence and outcome, as well as suggestions that certain therapies are preferable. As noted above, they can help to show that the results of clinical trials conducted in selected populations appear to apply in broader groups. At the present time, however, it is no substitute for a randomized clinical trial in evaluating whether one intervention is truly better than another.

Cross-Over Designs

The cross-over design is a special case of a randomized control trial and has some appeal to medical researchers. The cross-over design allows each participant to serve as his own control. In the simplest case, namely the two period cross-over design, each participant will receive either intervention or control (A or B) in the first period and the alternative in the succeeding period. The order in which A and B are given to each participant is randomized. Thus, approximately half of the participants receive the intervention in the sequence AB and the other half in the sequence BA. This is so that any trend from first period to second period can be eliminated in the estimate of group differences in response. Depending on the duration of expected action of the intervention (for example, drug half-life), a wash-out period may be used between the periods.

James et al. described 59 cross-over studies of analgesic agents. They concluded that if the studies had been designed using parallel or noncross-over designs, 2.4 times as many participants would have been needed [79]. Carriere showed that a three-period cross-over design is even more efficient than a two-period cross-over design [80]. A cross-over study need not have only two groups. A cross-over design for two active interventions and one control has been described [81].

The advantages and disadvantages of the two-period cross-over design have been described [19, 21, 81–84]. The appeal of the cross-over design to investigators

is that it allows assessment of how each participant does on both A and B. Since each participant is used twice, variability is reduced because the measured effect of the intervention is the difference in an individual participant's response to intervention and control. This reduction in variability enables investigators to use smaller sample sizes to detect a specific difference in response.

To use the cross-over design, however, a fairly strict assumption must be made; the effects of the intervention during the first period must not carry over into the second period. This assumption should be independent of which intervention was assigned during the first period and of the participant response. In many clinical trials, such an assumption is clearly inappropriate, even if a wash-out is incorporated. If, for example, the intervention during the first period cures the disease, then the participant obviously cannot return to the initial state. In other clinical trials, the cross-over design appears more reasonable. If a drug's effect is to lower blood pressure or heart rate, then a drug-versus-placebo cross-over design might be considered if the drug has no carryover effect once the participant is taken off medication. Obviously, a fatal event cannot serve as the primary response variable in a cross-over trial.

As indicated in the International Conference on Harmonisation document E9, Statistical Principles for Clinical Trials [85], cross-over trials should be limited to those situations with few losses of study participants. A typical and acceptable cross-over trial, for example, might compare two formulations of the same drug to assess bioequivalence in healthy participants. Similarly, different doses may be used to assess pharmacologic properties. In studies involving participants who are ill or otherwise have conditions likely to change, however, cross-over trials have the limitations noted above.

Although the statistical method for checking the assumption of no period–treatment interaction was described by Grizzle [86], the test is not as powerful as one would like. What decreases the power of the test is that the mean response of the AB group is compared to the mean response of the BA group. However, participant variability is introduced in this comparison, which inflates the error term in the statistical test. Thus, the ability to test the assumption of no period–intervention interaction is not sensitive enough to detect important violations of the assumption unless many participants are used. The basic appeal of the cross-over design is to avoid between-participant variation in estimating the intervention effect, thereby requiring a smaller sample size. Yet, the ability to justify the use of the design still depends on a test for carryover that includes between-participant variability. This weakens the main rationale for the cross-over design. Because of this insensitivity, the cross-over design is not as attractive as it at first appears. Fleiss et al. noted that even adjusting for baseline variables may not be adequate if inadequate time has been allowed to return to baseline at the start of the second period [87]. Brown [19, 21] and Hills and Armitage [88] discourage the use of the cross-over design in general. Only if there is substantial evidence that the therapy has no carryover effects, and the scientific community is convinced by that evidence, should a cross-over design be considered.

Withdrawal Studies

A number of studies have been conducted in which the participants on a particular treatment for a chronic disease are taken off therapy or have the dosage reduced. The objective is to assess response to discontinuation or reduction. This design may be validly used to evaluate the duration of benefit of an intervention already known to be useful. For example, subsequent to the Hypertension Detection and Follow-up Program [89], which demonstrated the benefits of treating mild and moderate hypertension, several investigators withdrew a sample of participants with controlled blood pressure from antihypertensive therapy [90]. Participants were randomly assigned to continue medication, stop medication yet initiate nutritional changes, or stop medication without nutritional changes. After 4 years, only 5% of those taken off medication without nutritional changes remained normotensive and did not need the re-instatement of medication. This compared with 39% who were taken off medication yet instituted weight loss and salt intake reductions.

Withdrawal studies have also been used to assess the efficacy of an intervention that had not conclusively been shown to be beneficial in the long term. An example is the Sixty Plus Reinfarction Study [91]. Participants doing well on oral anticoagulant therapy since their myocardial infarction, an average of 6 years earlier, were randomly assigned to continue on anticoagulants or assigned to placebo. Those who stayed on the intervention had lower mortality (not statistically significant) and a clear reduction in nonfatal reinfarction. A meta-analysis of prednisone and cyclosporine withdrawal trials (including some trials comparing withdrawal of the two drugs) in renal transplant patients has been conducted with graft failure or rejection as the response variables [92]. This meta-analysis found that withdrawal of prednisone was associated with increased risks of acute rejection and graft failure. Cyclosporine withdrawal led to an increase in acute rejection but not graft failure. The Fracture Intervention Trial Long-term Extension (FLEX) assessed the benefits of continuing treatment with alendronate after 5 years of therapy [93]. The group that was randomized to discontinue alendronate had a modest increase in vertebral fractures but no increase in nonvertebral fractures.

One serious limitation of this type of study is that a highly selected sample is evaluated. Only those participants who physicians thought were benefiting from the intervention were likely to have been on it for several months or years. Anyone who had major adverse effects from the drug would have been taken off and, therefore, not been eligible for the withdrawal study. Thus, this design can overestimate benefit and underestimate toxicity. Another drawback is that both participants and disease states change over time.

If withdrawal studies are conducted, the same standards should be adhered to that are used with other designs. Randomization, blinding where feasible, unbiased assessment, and proper data analysis are as important here as in other settings.

Factorial Design

In the simple case, the factorial design attempts to evaluate two interventions compared to control in a single experiment [2–4, 94]. See Table 5.1.

Given the cost and effort in recruiting participants and conducting clinical trials, getting two experiments done at once is appealing. Examples of factorial designs are the Canadian transient ischemic attack study where aspirin and sulfinpyrazone were compared with placebo [95], the Third International Study of Infarct Survival (ISIS-3) [96], the Physicians' Health Study [97], and the Women's Health Initiative (WHI) trial of hormone replacement, diet, and vitamin D plus calcium [98]. A review of analysis and reporting of factorial design trials [99] contains a list of 29 trials involving myocardial infarction and 15 other trials. Some factorial design studies are more complex than the 2 by 2 design, employing a third, or even a fourth level. It is also possible to leave some of the cells empty, that is, use an incomplete factorial design [100]. This was done in the Action to Control Cardiovascular Risk in Diabetes (ACCORD), which looked at intensive vs. less intensive glucose control plus either intensive blood pressure or lipid control [101]. This kind of design would be implemented if it is inappropriate, infeasible, or unethical to address every possible treatment combination. It is also possible to use a factorial design in a cross-over study [102].

The appeal of the factorial design might suggest that there really is a "free lunch." However, every design has strengths and weaknesses. A concern with the factorial design is the possibility of the existence of interaction and its impact on the sample size. Interaction means that the effect of intervention X differs depending upon the presence or absence of intervention Y, or vice versa. It is more likely to occur when the two drugs are expected to have related mechanisms of action.

If one could safely assume there were no interactions, one can show that with a modest increase in sample size, two experiments can be conducted in one; one which is considerably smaller than the sum of two independent trials under the same design specifications. However, if one cannot reasonably rule out interaction, one should statistically test for its presence. As is true for the cross-over design, the power for testing for interaction is less than the power for testing for

Table 5.1 Two-by-two factorial design

	Intervention X	Control	Marginals
Intervention Y	a	b	a + b
Control	c	d	c + d
Marginals	a + c	b + d	

Cell	*Intervention*
a	X + Y
b	Y + control
c	X + control
d	control + control

Effect of intervention X: a + c vs. b + d
Effect of intervention Y: a + b vs. c + d

the main effects of interventions (cells a + c vs. b + d or cells a + b vs. c + d). Thus, to obtain satisfactory power to detect interaction, the total sample size must be increased. The extent of the increase depends on the degree of interaction, which may not be known until the end of the trial. The larger the interaction, the smaller the increase in sample size needed to detect it. If an interaction is detected, or perhaps only suggested, the comparison of intervention X would have to be done individually for intervention Y and its control (cell a vs. b and cell c vs. d). The power for these comparisons is obviously less than for the a + c vs. b + d comparison.

As noted, in studies where the various interventions either act on the same response variable or possibly through the same mechanism of action, as with the presumed effect on platelets of both drugs in the Canadian transient ischemic attack study [95], interaction can be more of a concern. Furthermore, there may be a limited amount of reduction in the response variable that can be reasonably expected, restricting the joint effect of the interventions.

In trials such as the Physicians' Health Study [97], the two interventions, aspirin and beta carotene, were expected to act on two separate outcomes, cardiovascular disease and cancer. Thus, interaction was much less likely. But beta carotene is an antioxidant, and therefore might have affected both cancer and heart disease. It turned out to have no effect on either. Similarly, in the WHI [98], dietary and hormonal interventions may affect more than one disease process. There, diet had little effect on cancer and heart disease, but hormonal therapy had effects on heart disease, stroke, and cancer, among other conditions [103, 104].

In circumstances where there are two separate outcomes, e.g., heart disease and cancer, but one of the interventions may have an effect on both, data monitoring may become complicated. If, during the course of monitoring response variables it is determined that an intervention has a significant or important effect on one of the outcomes in a factorial design study, it may be difficult, or even impossible, to continue the trial to assess fully the effect on the other outcome. Chapter 16 reviews data monitoring in more detail.

The factorial design has some distinct advantages. If the interaction of two interventions is important to determine, or if there is little chance of interaction, then such a design with appropriate sample size can be very informative and efficient. However, the added complexity, impact on recruitment and adherence, and potential adverse effects of "polypharmacy" must be considered. Brittain and Wittes [105] discuss a number of settings in which factorial designs might be useful or not, and raise several cautions. In addition to the issue of interaction, they note that less than full adherence to the intervention can exacerbate problems in a factorial design trial.

Group Allocation Designs

In group or cluster allocation designs, a group of individuals, a clinic or a community are randomized to a particular intervention or control [106–110]. The rationale is that the intervention is most appropriately or more feasibly administered to an entire

group (for example, if the intervention consists of a broad media campaign). This design may also be better if there is concern about contamination. That is, when what one individual does might readily influence what other participants do. In the Child and Adolescent Trial for Cardiovascular Health, schools were randomized to different interventions [111]. A trial of vitamin A vs. placebo on morbidity and mortality in children in India randomized villages [112]. The Rapid Early Action for Coronary Treatment (REACT) trial involved ten matched pairs of cities. Within each pair, one city was randomly allocated to community education efforts aimed at reducing the time between symptoms of myocardial infarction and arrival at hospital [113]. Despite 18 months of community education, delay time was not different from that in the control cities. Communities have been compared in other trials [114, 115]. These designs have been used in cancer trials where a clinic or physician may have difficulty approaching people about the idea of randomization. The use of such designs in infectious disease control in areas with high prevalence of conditions such as tuberculosis and AIDS has become more common [116]. It should be noted that this example is both a group allocation design and a factorial design. In the group allocation design, the basic sampling units and the units of analysis are groups, not individual participants. This means that the effective sample is less than the total number of participants. Chapters 8 and 17 contain further discussions of the sample size determination and analysis of this design.

Hybrid Designs

Pocock [117] has argued that if a substantial amount of data is available from historical controls, then a hybrid, or combination design could be considered. Rather than a 50/50 allocation of participants, a smaller proportion could be randomized to control, permitting most to be assigned to the new intervention. A number of criteria must be met in order to combine the historical and randomized controls. These include the same entry criteria and evaluation factors and participant recruitment by the same clinic or investigator. The data from the historical control participants must also be fairly recent. This approach, if feasible, requires fewer participants to be entered into a trial. Machin, however, cautions that if biases introduced from the nonrandomized participants (historical controls) are substantial, more participants might have to be randomized to compensate than would be the case in a corresponding fully randomized trial [118].

Large, Simple and Pragmatic Clinical Trials

Advocates of large, simple trials maintain that for common pathological conditions, it is important to uncover even modest benefits of intervention, particularly short-term interventions that are easily implemented in a large population. They also

argue that an intervention is unlikely to have very different effects in different sorts of participants. Therefore, careful characterization of people at entry, or of interim response variables, is unnecessary. The important criteria for a valid study are unbiased (i.e., randomized) allocation of participants to intervention or control and unbiased assessment of outcome. Sufficiently large number of participants are more important than modest improvements in quality of data. The simplification of the study design and management allows for sufficiently large trials at reasonable cost. Examples of successful large, simple trials are ISIS [96], Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI) [119], Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries (GUSTO) [120], a study of digitalis [121], and the MICHELANGELO Organization to Assess Strategies in Acute Ischemic Syndromes (OASIS)-5 [122]. It should be noted that with the exception of the digitalis trial, these studies were relatively short term.

The questions addressed by these trials may be not only of the sort, "What treatment works better?" but also "What is the best way of providing the treatment?" Can something shown to work in an academic setting be translated to a typical community medical care setting? Several have advocated conducting pragmatic or practical clinical trials. These kinds of trials, as noted in Chap. 3, are conducted in clinical practices, often far from academic centers. They address questions perceived as relevant to those practices [123–125]. Because of the broad involvement of many practitioners, the results of the trial may be more widely applied than the results of a trial done in just major medical settings.

As indicated, these models depend upon a relatively easily administered intervention and an easily ascertained outcome. If the intervention is complex, requiring either special expertise or effort, particularly where adherence to protocol must be maintained over a long time, these kinds of studies are less likely to be successful. Similarly, if the response variable is a measure of morbidity that requires careful measurement by highly trained investigators, large simple or pragmatic trials are not feasible.

It has also been pointed out that baseline characteristics may be useful, not only for natural history studies but also for subgroup analysis. The issue of subgroup analysis is discussed more fully in Chap. 17. Although in general, it is likely that the effect of an intervention is qualitatively the same across subgroups, exceptions may exist. In addition, important quantitative differences may occur. When there is reasonable expectation of such differences, appropriate baseline variables need to be measured. Variables, such as age, gender, past history of a particular condition, or type of medication currently being taken, can be assessed in a simple trial. On the other hand, if an invasive laboratory test or a measurement that requires special training is necessary at baseline, such characterization may make a simple or pragmatic trial infeasible.

The investigator also needs to consider that the results of the trial must be persuasive to others. If other researchers or clinicians seriously question the validity of the trial because of inadequate information about participants or inadequate documentation of quality control, then the study has not achieved its purpose.

There is no doubt that many clinical trials are too expensive and too cumbersome, especially multicenter ones. The advent of the large, simple trial or the pragmatic trial

is an important step in enabling many meaningful medical questions to be addressed in an efficient manner. In other instances, however, the use of large number of participants may not compensate for reduced data collection and quality control. As always, the primary question being asked dictates the optimal design of the trial.

Studies of Equivalency and Noninferiority

Many clinical trials are designed to demonstrate that a new intervention is better than or superior to the control. However, not all trials have this goal. New interventions may have little or no superiority to existing therapies, but, as long as they are not materially worse, may be of interest because they are less toxic, less invasive, less costly, require fewer doses, improve quality of life, or have some other value to patients. In this setting, the goal of the trial would be to demonstrate that the new intervention is not worse, in terms of the primary response variable, than the standard by some predefined margin.

In studies of equivalency, the objective is to test whether a new intervention is equivalent to an established one. Noninferiority trials test whether the new intervention is no worse than, or at least as good as, some established intervention. Sample size issues for these kinds of trials are discussed in Chap. 8. In equivalency and noninferiority trials, several design aspects need to be considered [126–130]. The control or standard treatment must have been shown conclusively to be effective; that is, truly better than placebo or no therapy. The circumstances under which the active control was found to be useful ought to be reasonably close to those of the planned trial. Similarity of populations, concomitant therapy, and dosage are important. These requirements also mean that the trials that demonstrated efficacy of the standard should be recent and properly designed, conducted, analyzed, and reported. Table 5.2 shows the key assumptions for these trials.

First, the active control that is selected must be one that is an established standard for the indication being studied and not a therapy that is inferior to other known ones. It must be used with the dose and formulation proven effective. Second, the studies that demonstrated benefit of the control against either placebo or no treatment must be sufficiently recent such that no important medical advances or other changes have occurred, and in populations similar to those planned for the new trial. Third, the evidence that demonstrated the benefits of the control must be available so that a control group event rate can be estimated. Fourth, the response variable used in the new trial must be sensitive to the postulated effects of the control and intervention. The proposed trial must be able to demonstrate "assay sensitivity," or

Table 5.2 Noninferiority design assumptions

Proper control arm
Constancy over time and among participants
Availability of data from prior studies of the control
Assay sensitivity to demonstrate a true difference

the ability to show a difference if one truly exists. As emphasized in Chap. 8, the investigator must specify what she means by equivalence.

It cannot be shown statistically that two therapies are identical, as an infinite sample size would be required. Therefore, if the intervention falls sufficiently close to the standard, as defined by reasonable boundaries, the intervention is claimed to be "the same" as the control (in an equivalence trial) or no worse than the control (in a noninferiority trial). Selecting the margin of indifference or noninferiority, δ, is a challenge. Ideally, the relative risk of the new intervention compared to the control should be as close to 1 as possible. For practical reasons, the relative risk is often set in the range of 1.2–1.4. This means that in the worst case, the new intervention may be 20–40% inferior to standard treatment and yet be considered equivalent or noninferior. Some have even suggested that any new intervention could be approved by regulatory agencies as being noninferior to a standard control intervention if it retains as least 50% of the control vs. placebo effect. Further, there are options as to what 50% (or 40% or 20%) means. For example, one could choose either the point estimate from the control vs. placebo comparison, or the lower confidence interval estimate of that comparison. Also, the choice of the metric or scale must be selected, such as a relative risk, or hazard ratio, or perhaps an absolute difference. Of course, if an absolute difference that might seem reasonable with a high control group event rate is chosen, it might not seem so reasonable if the control group event rate turns out to be much lower than expected. This happened with a trial comparing warfarin against a new anticoagulant agent, where the observed control group event rate was less than that originally expected. Thus, with a predetermined absolute difference for noninferiority, the relative margin of noninferiority was larger than had been anticipated when the trial was designed [131].

It should be emphasized that new interventions are often hailed as successes if they are shown to be 20 or 25% better than placebo or a standard therapy. To turn around and claim that anything within a margin of 40 or 50% is equivalent to or noninferior to a standard therapy would seem illogical. But the impact on sample size of seeking to demonstrate that a new intervention is at most 20% worse than a standard therapy, rather than 40%, is considerable. As is discussed in Chap. 8, it would not be just a twofold increase in sample size, but a fourfold increase if the other parameters remained the same. Therefore, all design considerations and implications must be carefully considered.

Perhaps even more than in superiority trials, the quality, the size, and power of the new trial, and how well the trial is conducted, including how well participants comply with the assigned therapy, are crucial. A small sample size or poor compliance with the protocol, leading to low power, and therefore lack of significant difference, does not imply "equivalence."

To illustrate the concepts around noninferiority designs, consider the series of trials represented in Fig. 5.5, which depicts estimates with 95% confidence intervals for the intervention effect.

The heavy vertical line (labeled Delta) indicates the amount of worse effect of the intervention compared to the control that was chosen as tolerable. The thin vertical line indicates zero difference (a relative risk of 1). Trial A shows a new

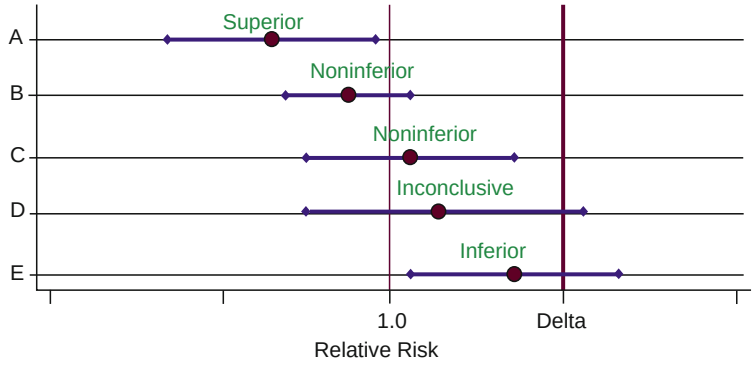


Fig. 5.5 Possible results of noninferiority trials; modified from [132]

intervention that is superior to control (i.e., the upper confidence interval excludes zero difference). Trial B has an estimate of the intervention effect that is favorable but the upper limit of the confidence interval does not exclude zero. It is less than the margin of indifference, however, and thus meets the criterion of being noninferior. Trial C is also noninferior, but the point estimate of the effect is slightly in favor of the control. Trial D does not conclusively show superiority or noninferiority, probably because it is too small or there were other factors that led to low power. Trial E indicates inferiority for the new intervention.

As discussed above, the investigator must consider several issues when designing an equivalence or noninferiority trial. First, the constancy assumption that the control vs. placebo effect has not changed over time is often not correct. This can be seen, for example, in two trials of the same design conducted back to back with essentially the same protocol and investigators, the PRAISE-1 and PRAISE-2 trials discussed in the section on Historical Controls and Databases [55, 56]. In PRAISE-1, the trial was stratified according to etiology, ischemic and nonischemic heart failure. Most of the favorable effect of the drug on mortality was seen in the nonischemic stratum, contrary to expectation. To validate that subgroup result, PRAISE-2 was conducted, using the same design, in nonischemic heart failure patients. In this second trial, no benefit of amlodipine was observed. The comparison of the placebo arms from PRAISE-1 and PRAISE-2 (Fig. 5.1) indicates that the two populations of nonischemic heart failure patients were at substantially different risk, despite being enrolled close in time, with the same entry criteria and same investigators. No covariate analysis could explain this difference in risk. Thus, the enrolled population itself is not constant, challenging the constancy assumption.

In addition, as background therapy changes, the effect of the control or placebo may also change. With more therapeutic options, the effect of one drug or intervention alone may no longer be as large as it was when placebo was the total background. Practice and referral patterns change.

Even if the data from prior trials of the selected control are available, the estimates of active control vs. placebo may not be completely accurate. As with all trials, effect

of treatment depends at least partly on the sample of participants who were identified and volunteered for the study. The observed effect is not likely to reflect the effect exactly in some other population. It is also possible that the quality of the trials used to obtain the effect of the control may not have been very good.

Many of the assumptions about the active control group event rates that go into the design of a noninferiority or equivalence trial are unlikely to be valid. At the end of the trial, investigators obtain seemingly more precise estimates of the margin and imputed "efficacy," when in fact they are based on a model that has considerable uncertainty and great care must be used in interpreting the results.

If I is the new intervention, C is the control or standard treatment, and P is placebo or no treatment, for the usual superiority trial, the goal is to show that the new intervention is better than placebo or no treatment, or that new intervention plus control is better than control alone.

$$I > P$$

$$I > C$$

$$I + C > C$$

For noninferiority trials, the margin of indifference, δ, is specified, where $I - C < \delta$. Efficacy imputation requires an estimate of the relative risk (RR) of the new intervention to control, RR(I/C) and of the control to placebo or no treatment, RR(C/P). Therefore, the estimated relative risk of the new intervention compared with placebo is

$$\mathrm{RR}(I/P) = \mathrm{RR}(I/C) \times \mathrm{RR}(C/P).$$

Rather than focus on the above assumption-filled model, an alternative approach might be considered. The first goal is to select the best control. This might be the one that, based on prior trials, was most effective. It might also be the one that the academic community considers as the standard of care, the one recommended in treatment guidelines, or the treatment that is most commonly used in practice. The selection will depend on the nature of the question being posed in the new trial. There might also be a choice of best controls, all considered to be similar, as, for example, one of several beta blockers or statins. The choice might be influenced by regulatory agencies. The margin of noninferiority should use the data from the prior trials of the active control to get some estimate for initiating discussion but should not use it as a precise value. Once that estimate has been obtained, investigators, with input from others, including, as appropriate, those from regulatory agencies, should use their experience and clinical judgment to make a final determination as to what margin of noninferiority would support using a new intervention. These decisions depend on factors such as the severity of the condition being studied, the known risks of the standard or control intervention, the trade-offs that might be achieved with the new intervention, whether it is 50 or 20%, or some other relative risk, or an absolute difference, and the practicality of obtaining the estimated

sample size. Having set the margin, effort must be on conducting the best trial, with as high participant adherence and complete follow-up as feasible. When the noninferiority trial has been completed, attention should be given to the interpretation of trial results, keeping in mind the entirety of the research using the new intervention and the active control and the relevance of the findings to the specific clinical practice setting (see Chaps. 17 and 19).

Adaptive Designs

There is a great deal of interest in designs, which are termed adaptive, but there are different designs that are adaptive and different meanings of the term. Clinical trials have used forms of adaptive designs for many years. As discussed in Chap. 1, early phase studies have designs that allow for modifications as the data accrue. Many late phase trials are adaptive in the sense that the protocol allows for modification of the intervention to achieve a certain goal, typically using an interim variable. For example, trials of antihypertensive agents, with the primary response variable of stroke or heart disease, will allow, and even encourage, changes in dose of the agent, or addition or substitution of agent to reach a specified blood pressure reduction or level. A trial in people with depression changed antidepression drugs based on interim success or lack of success as judged by depression questionnaires [133]. Some have proposed re-randomizing either all participants or those failing to respond adequately to the first drug to other agents [134, 135].

Some trials, by design, will adjust the sample size to retain a desired power if the overall event rate is lower than expected, the variability is higher than planned, or adherence is worse than expected. In such cases, the sample size can be recalculated using the updated information (see Chap. 8). An event-driven adaptive design continues until the number of events thought necessary to reach statistical significance, given the hypothesized intervention effect, accumulates. In trials where time to event is the outcome of interest, the length of follow-up or the number of study participants, or both, may be increased to obtain the predetermined number of outcome events. In other adaptive designs, the randomization ratio may be modified to keep the overall balance between intervention and control arms level on some risk score (see Chap. 6).

Various designs are called response adaptive. Traditionally, if the effect of the intervention was less than expected, or other factors led to a less than desirable conditional power, the study either continued to the end without providing a clear answer or was stopped early for futility (see Chap. 16). Some studies, particularly where the outcome occurred relatively quickly, allowed for modification of the randomization ratio between intervention and control arm, depending on the response of the most recent participant or responses of all accumulated participants.

Because of concerns about inefficiencies in study design, several trend adaptive approaches have been developed. At the beginning of the trial, the investigator may have inadequate information about the rate at which the outcome variable will occur

and be unable to make a realistic estimate of the effect of the intervention. Rather than continue to conduct an inappropriately powered trial or terminate early an otherwise well-designed study, the investigator may wish to modify the sample size. After a trial is underway and better estimates become available, these trend adaptive approaches adjust sample size based on the observed trend in the primary outcome, in order to maintain the desired power. Trend adaptive designs require some adjustment of the analysis to assess properly the significance of the test statistic. A criticism of these designs is that they can introduce bias during the implementation of the adjustment. They may also provide sufficient information to allow people not privy to the accumulating data to make reasonable guesses as to the trend.

Group sequential designs, in common use for many years, are also considered to be response adaptive in that they facilitate early termination of the trial when there is convincing evidence of benefit or harm. Response adaptive and trend adaptive designs will be considered further in Chaps. 16 and 17.

References

1. Fisher RA. *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd, 1925.
2. Fisher RA. *The Design of Experiments.* Edinburgh: Oliver and Boyd, 1935.
3. Cochran WG, Cox GM *Experimental Designs* (2nd edition). New York: John Wiley and Sons, 1957.
4. Cox DR. *Planning of Experiments.* New York: John Wiley and Sons, 1958.
5. Bull JP. The historical development of clinical therapeutic trials. *J Chronic Dis* 1959;10:218–248.
6. Eliot MM. The control of rickets: preliminary discussion of the demonstration in New Haven. *JAMA* 1925;85:656–663.
7. Hill AB. Observation and experiment. *N Engl J Med* 1953;248:995–1001.
8. Macfarlane G. *Howard Florey: The Making of a Great Scientist*. Oxford: Oxford University Press, 1979, pp 11–12.
9. Gocke DJ. Fulminant hepatitis treated with serum containing antibody to Australia antigen. *N Engl J Med* 1971;284:919.
10. Acute Hepatic Failure Study Group. Failure of specific immunotherapy in fulminant type B hepatitis. *Ann Intern Med* 1977;86:272–277.
11. Snow JB Jr, Kimmelman CP. Assessment of surgical procedures for Ménière's disease. *Laryngoscope* 1979;89:737–747.
12. Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research* (4th edition). Malden, MA: Blackwell Publishing, 2002.
13. Brown BW, Hollander M. *Statistics: A Biomedical Introduction.* New York: John Wiley and Sons, 1977.
14. Feinstein AR. *Clinical Biostatistics.* St Louis: The C.V. Mosby Company, 1977.
15. MacMahon B, Trichopoulos D. *Epidemiology: Principles and Methods* (2nd edition). Lippincott Williams & Wilkins, 1996.
16. Lilienfeld DE, Stolley PD. *Foundations of Epidemiology* (3rd edition). New York: Oxford University Press, 1994.
17. Srivastava JN (ed.). *A Survey of Statistical Design and Linear Models*. Amsterdam: North-Hollard, 1975.
18. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. 1. Introduction and design. *Br J Cancer* 1976;34:585–612.

19. Brown BW Jr. Statistical controversies in the design of clinical trials – some personal views. *Control Clin Trials* 1980;1:13–27.
20. Pocock SJ. Allocation of patients to treatment in clinical trials. *Biometrics* 1979;35:183–197.
21. Brown BW Jr. The crossover experiment for clinical trials. *Biometrics* 1980;36:69–79.
22. Hennekens CH, Buring JC. *Epidemiology in Medicine.* SL Mayrent (ed.). Boston: Little, Brown, 1987.
23. Byar DP. Some statistical considerations for design of cancer prevention trials. *Prev Med* 1989;18:688–699.
24. Geller NL (ed.). *Advances in Clinical Trial Biostatistics.* New York: Marcel Dekker, 2003.
25. Piantadosi S. *Clinical Trials: A Methodologic Perspective* (2nd edition). New York: John Wiley and Sons, 2005.
26. Machin D, Day S, Green S. *Textbook of Clinical Trials* (2nd edition). West Sussex: John Wiley and Sons, 2006.
27. Green S, Benedetti J, Crowley J. *Clinical Trials in Oncology* (2nd edition). Boca Raton: CRC Press, 2002.
28. Hulley SB, Cummings SR, Browner WS, et al. *Designing Clinical Research: An Epidemiologic Approach* (3rd edition). New York: Wolters Kluwer, 2006.
29. Meinert CL, *Clinical Trials: Design, Conduct and Analysis*. New York: Oxford University Press, 1986.
30. Cook TD, DeMets DL (eds.). *Introduction to Statistical Methods for Clinical Trials.* Boca Raton: Chapman & Hall/CRC, Taylor & Francis Group, LLC, 2008.
31. Chow S-C, Shao J. *Statistics in Drug Research: Methodologies and Recent Developments.* New York: Marcel Dekker, 2002.
32. Green SB, Byar DP. Using observational data from registries to compare treatments: the fallacy of omnimetrics. *Stat Med* 1984;3:361–373.
33. Gehan EA, Freireich EJ. Non-randomized controls in cancer clinical trials. *N Engl J Med* 1974;290:198–203.
34. Weinstein MC. Allocation of subjects in medical experiments. *N Engl J Med* 1974;291:1278–1285.
35. Byar DP, Simon RM, Friedewald WT, et al. Randomized clinical trials: perspectives on some recent ideas. *N Engl J Med* 1976;295:74–80.
36. Sapirstein W, Alpert S, Callahan TJ. The role of clinical trials in the Food and Drug Administration approval process for cardiovascular devices. *Circulation* 1994;89:1900–1902.
37. Hlatky MA. Perspective: evidence-based use of cardiac procedures and devices. *N Engl J Med* 2004;350:2126–2128.
38. Chalmers TC, Matta RJ, Smith H, Kunzier AM. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *N Engl J Med* 1977;297:1091–1096.
39. Peto R. Clinical trial methodology. *Biomedicine* (Special issue) 1978;28:24–36.
40. Goldman L, Feinstein AR. Anticoagulants and myocardial infarction: the problems of pooling, drowning, and floating. *Ann Intern Med* 1979;90:92–94.
41. Grace ND, Muench H, Chalmers TC. The present status of shunts for portal hypertension in cirrhosis. *Gastroenterology* 1966;50:684–691.
42. Sacks H, Chalmers TC, Smith H Jr. Randomized versus historical controls for clinical trials. *Am J Med* 1982;72:233–240.
43. Sacks HS, Chalmers TC, Smith H Jr. Sensitivity and specificity of clinical trials: randomized *v* historical controls. *Arch Intern Med* 1983;143:753–755.
44. Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 1983;309:1358–1361.
45. Ingelfinger FJ. The randomized clinical trial (editorial). *N Engl J Med* 1972;287:100–101.
46. Zelen M. A new design for randomized clinical trials. *N Engl J Med* 1979;300:1242–1245.
47. Anbar D. The relative efficiency of Zelen's prerandomization design for clinical trials. *Biometrics* 1983;39:711–718.
48. Ellenberg SS. Randomization designs in comparative clinical trials. *N Engl J Med* 1984;310: 1404–1408.

49. Zelen M. Randomized consent designs for clinical trials: an update. *Stat Med* 1990;9: 645–656.
50. Gehan EA. The evaluation of therapies: historical control studies. *Stat Med* 1984;3:315–324.
51. Lasagna L. Historical controls: the practitioner's clinical trials. *N Engl J Med* 1982;307: 1339–1340.
52. Moertel CG. Improving the efficiency of clinical trials: a medical perspective. *Stat Med* 1984;3:455–465.
53. Pocock SJ. Letter to the editor. *Br Med J* 1977;1:1661.
54. Veterans Administration Cooperative Urological Research Group. Treatment and survival of patients with cancer of the prostate. *Surg Gynecol Obstet* 1967;124:1011–1017.
55. Packer M, O'Connor CM, Ghali JK, et al. for the Prospective Randomized Amlodipine Survival Evaluation Study Group. Effect of amlodipine on morbidity and mortality in severe chronic heart failure. *N Engl J Med* 1996;335:1107–1114.
56. Thackray S, Witte K, Clark AL, Cleland JGF. Clinical trials update: OPTIME-CHF, PRAISE-2, ALLHAT. *Eur J Heart Fail* 2000;2:209–212.
57. Havlik RJ, Feinleib M (eds.). *Proceedings of the Conference on the Decline in Coronary Heart Disease Mortality.* Washington, D.C.: NIH Publication No. 79–1610, 1979.
58. U.S. Department of Health and Human Services. Health, United States, 2008, With Special Feature on the Health of Young Adults. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. http://www.cdc.gov/nchs/data/hus/hus08.pdf
59. Coronary Drug Project Research Group. Clofibrate and niacin in coronary heart disease. *JAMA* 1975;231:360–381.
60. Castro KG, Ward JW, Slutsker L, et al. 1993 revised classification system for HIV infection and expanded surveillance case definitions for AIDS among adolescents and adults. *MMWR Recomm Rep,* December 18, 1992;41(RR–17).
61. Current trends update: trends in AIDS diagnosis and reporting under the expanded surveillance definition for adolescents and adults – United States, 1993. *MMWR Weekly* 1994;43:826–831.
62. Rosenberg HM, Klebba AJ. Trends in cardiovascular mortality with a focus on ischemic heart disease: United States, 1950-1976. In Havlik R, Feinleib M (eds.). *Proceedings of the Conference on the Decline in Coronary Heart Disease Mortality.* Washington, D.C.: NIH Publication No. 79-1610, 1979.
63. National Heart, Lung, and Blood Institute. Morbidity and Mortality Chartbook on Cardiovascular, Lung, and Blood Diseases. National Heart, Lung, and Blood Institute, U.S. Department of Health and Human Services, Public Health Service. May 1994.
64. Bailar JC III, Louis TA, Lavori PW, Polansky M. Studies without internal controls. *N Engl J Med* 1984;311:156–162.
65. Dustan HP, Schneckloth RE, Corcoran AC, Page IH. The effectiveness of long-term treatment of malignant hypertension. *Circulation* 1958;18:644–651.
66. Bjork S, Sannerstedt R, Angervall G, Hood B. Treatment and prognosis in malignant hypertension: clinical follow-up study of 93 patients on modern medical treatment. *Acta Med Scand* 1960;166:175–187.
67. Bjork S, Sannerstedt R, Falkheden T, Hood B. The effect of active drug treatment in severe hypertensive disease: an analysis of survival rates in 381 cases on combined treatment with various hypotensive agents. *Acta Med Scand* 1961;169:673–689.
68. Starmer CF, Lee KL, Harrell FE, Rosati RA. On the complexity of investigating chronic illness. *Biometrics* 1980;36:333–335.
69. Hlatky MA, Lee KL, Harrell FE Jr, et al. Tying clinical research to patient care by use of an observational database. *Stat Med* 1984;3:375–387.
70. Hlatky MA, Califf RM, Harrell FE Jr, et al. Clinical judgment and therapeutic decision making. *J Am Coll Cardiol* 1990;15:1–14.
71. Moon TE, Jones SE, Bonadonna G, et al. Using a database of protocol studies to evaluate therapy: a breast cancer example. *Stat Med* 1984;3:333–339.

72. Anderson C. Measuring what works in health care. *Science* 1994;263:1080–1082.
73. Klungel OH, Heckbert SR, Longstreth WT, et al. Antihypertensive drug therapies and the risk of ischemic stroke. *Arch Intern Med* 2001;161:37–43.
74. Graham DJ, Campen D, Hui R, et al. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *Lancet* 2005;365:475–481.
75. Byar DP. Why databases should not replace randomized clinical trials. *Biometrics* 1980;36:337–342.
76. Dambrosia JM, Ellenberg JH. Statistical considerations for a medical database. *Biometrics* 1980;36:323–332.
77. Sheldon TA. Please bypass the PORT. *Br Med J* 1994;309:142–143.
78. Mantel N. Cautions on the use of medical databases. *Stat Med* 1983;2:355–362.
79. James KE, Forrest WH Jr, Rose RL. Crossover and noncrossover designs in four-point parallel line analgesic assays. *Clin Pharmacol Ther* 1985;37:242–252.
80. Carriere KC. Crossover designs for clinical trials. *Stat Med* 1994;13:1063–1069.
81. Koch GG, Amara IA, Brown BW Jr, et al. A two-period crossover design for the comparison of two active treatments and placebo. *Stat Med* 1989;8:487–504.
82. Fleiss JL. A critique of recent research on the two treatment crossover design. *Control Clin Trials* 1989;10:237–243.
83. Woods JR, Williams JG, Tavel M. The two-period crossover design in medical research. *Ann Intern Med* 1989;110:560–566.
84. Louis TA, Lavori PW, Bailar JC III, Polansky M. Crossover and self-controlled designs in clinical research. *N Engl J Med* 1984;310:24–31.
85. International Conference on Harmonisation: E9 Statistical principles for clinical trials. http://www.fda.gov/downloads/RegulatoryInformation/Guidances/UCM129505.pdf.
86. Grizzle JE. The two period change-over design and its use in clinical trials. *Biometrics* 1965;21:467–480.
87. Fleiss JL, Wallenstein S, Rosenfeld R. Adjusting for baseline measurements in the two-period crossover study: a cautionary note. *Control Clin Trials* 1985;6:192–197.
88. Hills M, Armitage P. The two-period cross-over clinical trial. *Br J Clin Pharmacol* 1979;8:7–20.
89. Hypertension Detection and Follow-up Program Cooperative Group. Five-year findings of the Hypertension Detection and Follow-Up Program 1 Reduction in mortality of persons with high blood pressure, including mild hypertension. *JAMA* 1979;242:2562–2571.
90. Stamler R, Stamler J, Grimm R, et al. Nutritional therapy for high blood pressure-Final report of a four-year randomized controlled trial – The Hypertension Control Program. *JAMA* 1987;257:1484–1491.
91. Report of the Sixty Plus Reinfarction Study Research Group. A double-blind trial to assess long-term oral anticoagulant therapy in elderly patients after myocardial infarction. *Lancet* 1980;316:989–994.
92. Kasiske BL, Chakkera HA, Louis TA, Ma JZ. A meta-analysis of immunosuppression withdrawal trials in renal transplantation. *J Am Soc Nephrol* 2000;11:1910–1917.
93. Black DM, Schwartz AV, Ensrud KE, et al, for the FLEX Research Group. Effects of continuing or stopping alendronate after 5 years of treatment: The Fracture Intervention Trial Long-term Extension (FLEX): a randomized trial. *JAMA* 2006;296:2927–2938.
94. Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomized controlled trials. *BMC Med Res Methodol* 2003;3:26; doi:10.1186/1471-2288-3-26.
95. The Canadian Cooperative Study Group. A randomized trial of aspirin and sulfinpyrazone in threatened stroke. *N Engl J Med* 1978;299:53–59.
96. ISIS-3 (Third International Study of Infarct Survival) Collaborative Group. ISIS-3: a randomized study of streptokinase vs plasminogen activator vs anistrephase and of aspirin plus heparin vs aspirin alone among 41,299 cases of suspected acute myocardial infarction. *Lancet* 1992;339:753–770.

97. Stampfer MJ, Buring JE, Willett W, et al. The 2 x 2 factorial design: its application to a randomized trial of aspirin and carotene in U.S. physicians. *Stat Med* 1985;4:111–116.
98. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control Clin Trials* 1998;19:61–109.
99. McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. *JAMA* 2003;289:2545–2553.
100. Byar DP, Herzberg AM, Tan W-Y. Incomplete factorial designs for randomized clinical trials. *Stat Med* 1993;12:1629–1641.
101. Action to Control Cardiovascular Risk in Diabetes Study Group. Effects of intensive glucose lowering in Type 2 diabetes. *N Engl J Med* 2008;358:2545–2559.
102. Fletcher DJ, Lewis SM, Matthews JNS. Factorial designs for crossover clinical trials. *Stat Med* 1990;9:1121–1129.
103. Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *JAMA* 2002;288:321–333.
104. The Women's Health Initiative Steering Committee. Effects of conjugated equine estrogen in postmenopausal women with hysterectomy. *JAMA* 2004;291:1701–1712.
105. Brittain E, Wittes J. Factorial designs in clinical trials: the effects of non-compliance and subadditivity. *Stat Med* 1989;8:161–171.
106. Hayes RJ, Moulton LH. *Cluster Randomized Trials: A Practical Approach.* Chapman & Hall/CRC, Taylor & Francis Group, 2009.
107. Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. *Am J Epidemiol* 1981;114:906–914.
108. Armitage P. The role of randomization in clinical trials. *Stat Med* 1982;1:345–352.
109. Simon R. Composite randomization designs for clinical trials. *Biometrics* 1981;37:723–731.
110. Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol* 1978;108:100–102.
111. Zucker DM, Lakatos E, Webber LS, et al. Statistical design of the Child and Adolescent Trial for Cardiovascular Health (CATCH): implications of cluster randomization. *Control Clin Trials* 1995;16:96–118.
112. Vijayaraghavan K, Radhaiah G, Prakasam BS, et al. Effect of massive dose vitamin A on morbidity and mortality in Indian children. *Lancet* 1990;336:1342–1345.
113. Luepker RV, Raczynski JM, Osganian S, et al. Effect of a community intervention on patient delay and emergency medical service use in acute coronary heart disease: the Rapid Early Action for Coronary Treatment (REACT) trial. *JAMA* 2000;284:60–67.
114. Farquhar JW, Fortmann SP, Flora JA, et al. Effects of community-wide education on cardiovascular disease risk factors. The Stanford Five-City Project. *JAMA* 1990;264:359–365.
115. Gail MH, Byar DP, Pechacek TF, Corle DK, for COMMIT Study Group. Aspects of statistical design for the Community Intervention Trial for Smoking Cessation. *Control Clin Trials* 1992;13:6–21.
116. Sismanidis C, Moulton LH, Ayles H, et al. Restricted randomization of ZAMSTAR: a 2x2 factorial cluster randomized trial. *Clin Trials* 2008;5:316–327.
117. Pocock SJ. The combination of randomized and historical controls in clinical trials. *J Chronic Dis* 1976;29:175–188.
118. Machin D. On the possibility of incorporating patients from nonrandomising centres into a randomised clinical trial. *J Chronic Dis* 1979;32:347–353.
119. Gruppo Italiano per lo Studio della Streptochinasi nell' Infarto Miocardico (GISSI). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. *Lancet* 1986;i:397–402.
120. The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Engl J Med* 1993;329:673–682; correction *N Engl J Med* 1994;331:277.
121. The Digitalis Investigation Group. Rationale, design, implementation, and baseline characteristics of patients in the DIG Trial: a large, simple, long-term trial to evaluate the effect of digitalis on mortality in heart failure. *Control Clin Trials* 1996;17:77–97.

122. MICHELANGELO OASIS 5 Steering Committee. Design and rationale of the MICHELANGELO Organization to Assess Strategies in Acute Ischemic Syndromes (OASIS)-5 trial program evaluating fondaparinux, a synthetic factor Xa inhibitor, in patients with non-ST-segment elevation acute coronary syndromes. *Am Heart J* 2005;150:1107–1114.
123. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA* 2003;290:1624–1632.
124. March JS, Silva SG, Compton S, et al. The case for practical clinical trials in psychiatry. *Am J Psychiatry* 2005;162:836–846.
125. Thorpe KE, Zwarenstein M, Oxman AD, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *J Clin Epidemiol* 2009;62:464–475.
126. Blackwelder WC. "Proving the null hypothesis" in clinical trials. *Control Clin Trials* 1982;3: 345–353.
127. Hung JHM, Wang SJ, Tsong Y, et al. Some fundamental issues with non-inferiority testing in active controlled trials. *Stat Med* 2003;30:213–225.
128. Fleming TR. Current issues in non-inferiority trials. *Stat Med* 2008;27:317–332.
129. D'Agostino RB Sr, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics. *Stat Med* 2003;22:169–186.
130. Kaul S, Diamond GA. Making sense of noninferiority: a clinical and statistical perspective on its application to cardiovascular clinical trials. *Prog Cardiovasc Dis* 2007;49:284–299.
131. SPORTIF Executive Steering Committee for the SPORTIF V Investigators. Ximelagatran vs warfarin for stroke prevention in patients with nonvalvular atrial fibrillation: a randomized trial. *JAMA* 2005;293:690–698.
132. Pocock SJ, Ware JH. Translating statistical findings into plain English. *Lancet* 2009;373:1926–1928.
133. Trivedi MH, Rush AJ, Wisniewski SR, et al, for the STAR*D Study Team. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *Am J Psychiatry* 2006;163:28–40.
134. Murphy SA, Oslin DW, Rush AJ, Zhu J, for MCATS. Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders (Perspective). *Neuropsychopharmacology* 2007;32:257–262.
135. Lavori PW, Dawson D. Improving the efficiency of estimation in randomized trials of adaptive treatment strategies. *Clin Trials* 2007;4:297–308.

Chapter 6
The Randomization Process

The randomized controlled clinical trial is the standard by which all trials are judged since other designs have certain undesirable features. In the simplest case, randomization is a process by which each participant has the same chance of being assigned to either intervention or control. An example would be the toss of a coin, in which heads indicates intervention group and tails indicates control group. Even in the more complex randomization strategies, the element of chance underlies the allocation process. Of course, neither trial participant nor investigator should know what the assignment will be before the participant's decision to enter the study. Otherwise, the benefits of randomization can be lost. The role that randomization plays in clinical trials has been discussed in Chap. 5 as well as by numerous authors [1–12]. While not all accept that randomization is essential [11, 12], most agree it is the best method for achieving comparability between study groups and is the basis for statistical inference [2, 3].

Fundamental Point

Randomization tends to produce study groups comparable with respect to known as well as unknown risk factors, removes investigator bias in the allocation of participants, and guarantees that statistical tests will have valid false positive error rates.

Several methods for randomly allocating participants are used [4, 5, 10, 13, 14]. This chapter presents the most common of these methods and considers the advantages and disadvantages of each. Unless stated otherwise, it can be assumed that the randomization strategy will allocate participants into two groups, an intervention group and a control group. However, many of the methods described here can easily be generalized for use with more than two groups.

Two forms of experimental bias are of concern. The first, *selection bias*, occurs if the allocation process is predictable [9, 15–18]. In this case, the decision to enter a participant into a trial may be influenced by the anticipated treatment assignment. If any bias exists as to what treatment particular types of participants should

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_6, © Springer Science+Business Media, LLC 2010

receive, then a selection bias might occur. All of the randomization procedures described avoid selection bias by not being predictable. A second bias, *accidental bias*, can arise if the randomization procedure does not achieve balance on risk factors or prognostic covariates. Some of the allocation procedures described are more vulnerable to accidental bias, especially for small studies. For large studies, however, the chance of accidental bias is negligible [9].

Whatever randomization process is used, the report of the trial should contain a brief, but clear description of that method. In the 1980s, Altman and Doré [17] reported a survey of four medical journals where 30% of published randomized trials gave no evidence that randomization had in fact been used. As many as 10% of these "randomized" trials in fact used nonrandom allocation procedures. Sixty percent did not report the type of randomization that was used. In one review in the 1990s, only 20–30% of trials provided fair or adequate descriptions, depending on the size of the trial or whether the trial was single center or multicenter [18]. More recently, a review of 253 trials published in five major medical journals after the release of the Consolidated Standards for Reporting Trials (CONSORT) [19] recommendations found little improvement in reports of how randomization was accomplished [20]. Descriptions need not be lengthy to inform the reader, publications should clearly indicate the type of randomization method and how the randomization was implemented.

Fixed Allocation Randomization

Fixed allocation procedures assign the interventions to participants with a prespecified probability, usually equal, and that allocation probability is not altered as the study progresses. A number of methods exist by which fixed allocation is achieved [4, 5, 10, 14, 21–25], and we review three of these – simple, blocked, and stratified.

Our view is that allocation to intervention and control groups should be equal unless there are compelling reasons to do otherwise. Peto [6], among others, has suggested an unequal allocation ratio, such as 2:1, of intervention to control. The rationale for such an allocation is that the study may slightly lose sensitivity but may gain more information about participant responses to the new intervention, such as toxicity and side effects. In some instances, less information may be needed about the control group and, therefore, fewer control participants are required. If the intervention turns out to be beneficial, more study participants would benefit than under an equal allocation scheme. However, new interventions may also turn out to be harmful, in which case more participants would receive them under the unequal allocation strategy. Although the loss of sensitivity or power may be less than 5% for allocation ratios approximately between 1/2 and 2/3 [7, 21], equal allocation is the most powerful design and therefore generally recommended. We also believe that equal allocation is more consistent with the view of indifference or equipoise toward which of the two groups a participant is assigned. Unequal allocation may

indicate to the participants and to their personal physicians that one intervention is preferred over the other. In a few circumstances, the cost of one treatment may be extreme so that an unequal allocation of 2:1 or 3:1 may help to contain costs while not causing a serious loss of power. Thus, there are trade-offs that must be considered. In general, equal allocation will be presumed throughout the following discussion unless otherwise indicated.

Simple Randomization

The most elementary form of randomization, referred to as simple or complete randomization, is best illustrated by a few examples [4, 5]. One simple method is to toss an unbiased coin each time a participant is eligible to be randomized. For example, if the coin turns up heads, the participant is assigned to group *A*; if tails, to group *B*. Using this procedure, approximately one half of the participants will be in group *A* and one half in group *B*. In practice, for small studies, instead of tossing a coin to generate a randomization schedule, a random digit table on which the equally likely digits 0–9 are arranged by rows and columns is usually used to accomplish simple randomization. By randomly selecting a certain row (column) and observing the sequence of digits in that row (column) *A* could be assigned, for example, to those participants for whom the next digit was even and *B* to those for whom the next digit was odd. This process produces a sequence of assignments, which is random in order, and each participant has an equal chance of being assigned to *A* or *B*.

For large studies, a more convenient method for producing a randomization schedule is to use a random number producing algorithm, available on most computer systems. A simple randomization procedure might assign participants to group *A* with probability p and participants to group *B* with probability $1-p$. One computerized process for simple randomization is to use a uniform random number algorithm to produce random numbers in the interval from 0.0 to 1.0. Using a uniform random number generator, a random number can be produced for each participant. If the random number is between 0 and p, the participant would be assigned to group *A*; otherwise to group *B*. For equal allocation, the probability cut point, p, is one-half (i.e., $p=0.50$). If equal allocation between *A* and *B* is not desired ($p \neq 1/2$), then p can be set to the desired proportion in the algorithm and the study will have, on the average, a proportion p of the participants in group *A*.

This procedure can be adapted easily to more than two groups. Suppose, for example, the trial has three groups, *A*, *B*, and *C*, and participants are to be randomized such that a participant has a 1/4 chance of being in group *A*, a 1/4 chance of being in group *B*, and a 1/2 chance of being in group *C*. By dividing the interval 0–1 into three pieces of length 1/4, 1/4, and 1/2, random numbers generated will have probabilities of 1/4, 1/4, and 1/2, respectively, of falling into each subinterval. Specifically, the intervals would be 0–0.249, 0.25–0.499, and 0.50–1.0. Then, any participant whose random number falls between 0 and 0.249 is assigned *A*, any

participant whose random number falls between 0.25 and 0.499 is assigned *B*, and the others, *C*. For equal allocation, the interval would be divided into thirds and assignments made accordingly.

The advantage of this simple randomization procedure is that it is easy to implement. The major disadvantage is that, although in the long run the number of participants in each group will be in the proportion anticipated, at any point in the randomization, including the end, there could be a substantial imbalance [22]. This is true particularly if the sample size is small. For example, if 20 participants are randomized with equal probability to two treatment groups, the chance of a 12:8 split (i.e., 60% *A*, 40% *B*) or worse is approximately 50%. For 100 participants, the chance of the same ratio (60:40 split) or worse is only 5%. While such imbalances do not cause the statistical tests to be invalid, they do reduce ability to detect true differences between the two groups. In addition, such imbalances appear awkward and may lead to some loss of credibility for the trial, especially for the person not oriented to statistics. For this reason primarily, simple randomization is not often used, even for large studies. In addition, interim analysis of accumulating data might be difficult to interpret with major imbalances in number of participants per arm, especially for smaller trials.

Some investigators incorrectly believe that an alternating assignment of participants to the intervention and the control groups (e.g., *ABABAB*…) is a form of randomization. However, no random component exists in this type of allocation except perhaps for the first participant. A major criticism of this method is that, in a single-blind or unblinded study, the investigators know the next assignment, which could lead to a bias in the selection of participants. Even in a double-blind study, if the blind is broken on one participant as sometimes happens, the entire sequence of assignments is known. Therefore, this type of allocation method should be avoided.

Blocked Randomization

Blocked randomization, sometimes called permuted block randomization, was described by Hill [1] in 1951. It is used in order to avoid serious imbalance in the number of participants assigned to each group – an imbalance that could occur in the simple randomization procedure. Blocked randomization guarantees that at no time during randomization will the imbalance be large and that at certain points the number of participants in each group will be equal [4, 5, 26].

If participants are randomly assigned with equal probability to groups *A* or *B*, then for each block of even size (e.g., 4, 6, or 8), one half of the participants will be assigned to *A* and the other half to *B*. The order in which the interventions are assigned in each block is randomized, and this process is repeated for consecutive blocks of participants until all participants are randomized. For example, the investigators may want to ensure that after every fourth randomized participant, the number of participants in each intervention group is equal. Then, a block of size 4

would be used, and the process would randomize the order in which two *A*s and two *B*s are assigned for every consecutive group of four participants entering the trial. One may write down all the ways of arranging the groups and then randomize the order in which these combinations are selected. In the case of block size 4, there are six possible combinations of group assignments: *AABB*, *ABAB*, *BAAB*, *BABA*, *BBAA*, and *ABBA*. One of these arrangements is selected at random and the four participants are assigned accordingly. This process is repeated as many times as needed.

Another method of blocked randomization may also be used. In this method for randomizing the order of assignments within a block of size *b*, a random number between 0 and 1 for each of the *b* assignments (half of which are *A* and the other half *B*) is obtained. The example below illustrates the procedure for a block of size four (two *A*s and two *B*s). Four random numbers are drawn between 0 and 1 in the order shown.

Assignment	Random number	Rank
A	0.069	1
A	0.734	3
B	0.867	4
B	0.312	2

The assignments then are ranked according to the size of the random numbers. This leads to the assignment order of *ABAB*. This process is repeated for another set of four participants until all have been randomized.

The advantage of blocking is that balance between the number of participants in each group is guaranteed during the course of randomization. The number in each group will never differ by more than $b/2$ when b is the length of the block. This can be important for at least two reasons. First, if the type of participant recruited for the study changes during the entry period, blocking will produce more comparable groups. For example, an investigator may use different sources of potential participants sequentially. Participants from these sources may vary in severity of illness or other crucial respects. One source, with the more seriously ill participants, may be used early during enrollment and another source, with healthier participants, late in enrollment [3]. If the randomization were not blocked, more of the seriously ill participants might be randomized to one group. Because the later participants are not as sick, this early imbalance would not be corrected. A second advantage of blocking is that if the trial should be terminated before enrollment is completed, balance will exist in terms of number of participants randomized to each group.

A potential, but solvable problem with basic blocked randomization is that if the blocking factor *b* is known by the study staff and the study is not double-blind, the assignment for the last person entered in each block is known before entry of that person. For example, if the blocking factor is 4 and the first three assignments are *ABB*, then the next assignment must be *A*. This could, of course, permit a bias in the selection of every fourth participant to be entered. Clearly, there is no reason to make the blocking factor known. However, in a study that is not double-blind, with

a little ingenuity the staff can soon discover the blocking factor. For this reason, repeated blocks of size 2 should not be used. On a few occasions, perhaps as an intellectual challenge, investigators or their clinic staff have attempted to break the randomization scheme. This curiosity is natural but nevertheless can cause problems in the integrity of the randomization process. To avoid this problem in the trial that is not double-blind, the blocking factor can be varied as the recruitment continues. In fact, after each block has been completed, the size of the next block could be determined in a random fashion from a few possibilities such as 2, 4, 6, and 8. The probabilities of selecting a block size can be set at whatever values one wishes with the constraint that their sum equals 1.0. For example, the probabilities of selecting block sizes 2, 4, 6, and 8 can be 1/6, 1/6, 1/3, and 1/3, respectively. Randomly selecting the block size makes it very difficult to determine where blocks start and stop and thus determine the next assignment.

A disadvantage of blocked randomization is that, from a strictly theoretical point of view, analysis of the data is more complicated than if simple randomization were used. The data analysis performed at the end of the study should reflect the randomization process actually performed [26–31]. This requirement would complicate the analysis because many analytical methods assume a simple randomization. In their analysis of the data most investigators ignore the fact that the randomization was blocked. Matts and McHugh [31] studied this problem and concluded that the measurement of variability used in the statistical analysis is not exactly correct if the blocking is ignored. Since blocking guarantees balance between the two groups and, therefore, increases the power of a study, blocked randomization with the appropriate analysis is more powerful than not blocking at all or blocking and then ignoring it in the analysis. Statisticians recognize the problem and feel that, at worst, they are being conservative by ignoring the fact that the randomization was blocked [14]. That is, the study will have probably slightly less power than it could have with the correct analysis, and the "true" significance level is more extreme than that computed.

Stratified Randomization

One of the objectives in allocating participants is to achieve between group comparability of certain characteristics known as prognostic or risk factors [4, 32–45]. Measured at baseline, these are factors that correlate with subsequent participant response or outcome. Investigators may become concerned when prognostic factors are not evenly distributed between intervention and control groups. As indicated previously, randomization tends to produce groups which are, on the average, similar in their entry characteristics, both known and unknown. This is a concept likely to be true for large studies or for many small studies when averaged. For any single study, especially a small study, there is no guarantee that all baseline characteristics will be similar in the two groups. In the multicenter Aspirin Myocardial Infarction Study [46], which had 4,524 participants, the top 20 cardiovascular prognostic factors for total mortality identified in the Coronary Drug Project [33] were compared

in the intervention and control groups, and no major differences were found (Furberg CD, unpublished data). However, individual clinics, with an average of 150 participants, showed considerable imbalance for many variables between the groups. Imbalances in prognostic factors can be dealt with either after the fact by using stratification in the analysis (Chap. 17) or can be prevented by using stratification in the randomization. Stratified randomization is a method that helps achieve comparability between the study groups for those factors considered.

Stratified randomization requires that the prognostic factors be measured either before or at the time of randomization. If a single factor is used, it is divided into two or more subgroups or strata (e.g., age 30–34 years, 35–39 years, 40–44 years). If several factors are used, a stratum is formed by selecting one subgroup from each of them. The total number of strata is the product of the number of subgroups in each factor. The stratified randomization process involves measuring the level of the selected factors for a participant, determining to which stratum she belongs and performing the randomization within that stratum.

Within each stratum, the randomization process itself could be simple randomization, but in practice most clinical trials use some blocked randomization strategy. Under a simple randomization process, imbalances in the number in each group within the stratum could easily happen and thus defeat the purpose of the stratification. Blocked randomization is, as described previously, a special kind of stratification. However, this text will restrict use of the term blocked randomization to stratifying over time, and use stratified randomization to refer to stratifying on factors other than time. Some confusion may arise here because early texts on design used the term blocking as this book uses the term stratifying. However, the definition herein is consistent with current usage in clinical trials.

As an example of stratified randomization with a block size of 4, suppose an investigator wants to stratify on age, sex, and smoking history. One possible classification of the factors would be three 10-year age levels and three smoking levels.

Age (years)	Sex	Smoking history
1. 40–49	1. Male	1. Current smoker
2. 50–59	2. Female	2. Ex-smoker
3. 60–69		3. Never smoked

Thus, the design has $3 \times 2 \times 3 = 18$ strata. The randomization for this example appears in Table 6.1.

Participants who were between 40 and 49 years old, male and current smokers, that is, in stratum 1, would be assigned to groups *A* or *B* in the sequences *ABBA BABA*… Similarly, random sequences would appear in the other strata.

Small studies are the ones most likely to require stratified randomization, because in large studies, the magnitude of the numbers increases the chance of comparability of the groups. In the example shown above, with three levels of the first factor (age), two levels of the second factor (sex), and three levels of the third factor (smoking history), 18 strata have been created. As factors are added and the levels within factors are refined, the number of strata increases rapidly. If the example with 18 strata had

Table 6.1 Stratified randomization with block size of four

Strata	Age	Sex	Smoking	Group assignment
1	40–49	M	Current	*ABBA BABA*…
2	40–49	M	Ex	*BABA BBAA*…
3	40–49	M	Never	etc.
4	40–49	F	Current	
5	40–49	F	Ex	
6	40–49	F	Never	
7	50–59	M	Current	
8	50–59	M	Ex	
9	50–59	M	Never	
10	50–59	F	Current	
11	50–59	F	Ex	
12	50–59	F	Never	
	(etc.)			

100 participants to be randomized, then only five to six participants would be expected per stratum if the study population were evenly distributed among the levels. Since the population is most likely not evenly distributed over the strata, some strata would actually get fewer than five to six participants. If the number of strata were increased, the number of participants in each stratum would be even fewer. Pocock and Simon [34] showed that increased stratification in small studies can be self-defeating because of the sparseness of data within each stratum. Thus, only important variables should be chosen and the number of strata kept to a minimum.

In addition to making the two study groups appear comparable with regard to specified factors, the power of the study can be increased by taking the stratification into account in the analysis. Stratified randomization, in a sense, breaks the trial down into smaller trials. Participants in each of the "smaller trials" belong to the same stratum. This reduces variability in group comparisons if the stratification is used in the analysis. Reduction in variability allows a study of a given size to detect smaller group differences in response variables or to detect a specified difference with fewer participants [25, 26].

Sometimes the variables initially thought to be most prognostic and, therefore used in the stratified randomization, turn out to be unimportant. Other factors may be identified later which, for the particular study, are of more importance. If randomization is done without stratification, then analysis can take into account those factors of interest and will not be complicated by factors thought to be important at the time of randomization. It has been argued that there usually does not exist a need to stratify at randomization because stratification at the time of analysis will achieve nearly the same expected power [6]. This issue of stratifying pre vs. post-randomization has been widely discussed [37–40, 43]. It appears for a large study that stratification after randomization provides nearly equal efficiency to stratification before randomization [44, 45]. However, for studies of 100 participants or fewer, stratifying the randomization using two or three prognostic factors may achieve greater power although the increase may not be large.

Stratified randomization is not the complete solution to all potential problems of baseline imbalance. Another strategy for small studies with many prognostic factors is considered below in the section on adaptive randomization.

In multicenter trials, centers vary with respect to the type of participants randomized as well as the quality and type of care given to participants during follow-up. Thus, the center may be an important factor related to participant outcome, and the randomization process should be stratified accordingly [41]. Each center then represents, in a sense, a replication of the trial, though the number of participants within a center is not adequate to answer the primary question. Nevertheless, results at individual centers can be compared to see if trends are consistent with overall results. Another reason for stratification by center is that if a center should have to leave the study, the balance in prognostic factors in other centers would not be affected.

One further point might need consideration. If in the stratified randomization, a specific proportion or quota is intended for each stratum, the recruitment of eligible participants might not be at the same rate. That is, one stratum might meet the target before the others. If a target proportion is intended, then plans need to be in place to close down recruitment for that stratum, allowing the others to be completed.

Adaptive Randomization Procedures

The randomization procedures described in the sections on fixed allocation above are nonadaptive strategies. In contrast, adaptive procedures change the allocation probabilities as the study progresses. Two types of adaptive procedures will be considered here. First, we will discuss methods which adjust or adapt the allocation probabilities according to imbalances in numbers of participants or in baseline characteristics between the two groups. Second, we will briefly review adaptive procedures that adjust allocation probabilities according to the responses of participants to the assigned intervention.

Baseline Adaptive Randomization Procedures

The Biased Coin Randomization procedure, originally discussed by Efron [47], attempts to balance the number of participants in each treatment group based on the previous assignments but does not take participant responses into consideration. Several variations to this approach have been discussed [48–64]. The purpose of the algorithm is basically to randomize the allocation of participants to groups *A* and *B* with equal probability as long as the number of participants in each group is equal or nearly equal. If an imbalance occurs and the difference in the number of participants is greater than some prespecified value, the allocation probability (p) is adjusted so that the probability is higher for the group with fewer participants.

The investigator can determine the value of the allocation probability he wishes to use. The larger the value of p, the faster the imbalance will be corrected, while the nearer p is to 0.5, the slower the correction. Efron suggests an allocation probability of $p=2/3$ when a correction is indicated. Since much of the time p is greater than 1/2, the process has been named the "biased coin" method. As a simple example, suppose n_A and n_B represent the number of participants in groups A and B, respectively. If n_A is less than n_B and the difference exceeds a predetermined value, D, then we allocate the next participant to group A with probability $p=2/3$. If n_A is greater than n_B by an amount of D, we allocate to group B with probability $p=2/3$. Otherwise, p is set at 0.50. This procedure can be modified to include consideration of the number of consecutive assignments to the same group and the length of such a run.

This approach, from a strictly theoretical point of view, demands a cumbersome data analysis process. The correct analysis requires that the significance level for the test statistic be determined by considering all possible sequences of assignments, which could have been made in repeated experiments using the same biased coin allocation rule where no group differences are assumed to exist. Although this is feasible to do with digital computers, the analysis is not easy. As with the blocked randomization scheme, the analysis often ignores this requirement. Efron [47] argues that it is probably not necessary to take the biased coin randomization into account in the analysis, especially for larger studies. However, a test statistic, which ignores the biased coin randomization will not provide the correct variance term. Most often, the variance will be larger than it would be with proper calculation, thus giving a conservative test in the sense that the probability of rejecting the null hypothesis is less than it would be if the proper analysis were used. One possible advantage of the biased coin approach over the blocked randomization scheme is that the investigator cannot determine the next assignment by discovering the blocking factor. However, the biased coin method does not appear to be as widely used as the blocked randomization scheme because of its complexity.

Another similar adaptive randomization method is referred to as the Urn Design, based on the work of Wei and colleagues [65–68]. This method also attempts to keep the number of participants randomized to each group reasonably balanced as the trial progresses. The name Urn Design refers to the conceptual process of randomization. Imagine an urn filled with m red balls and m black balls. If a red ball is drawn at random, assign the participant to group A, return the red ball, and add a black ball to the urn. If a black ball is drawn, assign the participant to group B, return that ball, and add a red ball to the urn. This process will keep the number of participants in each group reasonably close because it adjusts the allocation probability. From a theoretical point of view, this method, like the biased coin design, would require the analyses to account for the randomization [65]. While this is possible, these analyses are not straightforward. It seems likely, as for the biased coin design, that if this randomization method is used, but ignored in the analyses, the p-value will be slightly conservative, that is, slightly larger than if the strictly correct analysis were done. The urn model was used successfully in the multicenter Diabetes Control and Complication Trial [69].

Other stratification methods are adaptive in the sense that intervention assignment probabilities for a participant are a function of the distribution of the prognostic factors for participants already randomized. This concept was suggested by Efron [47] as an extension of the biased coin method and also has been discussed in depth by Pocock and Simon [34] and others [48–53, 70–72]. In a simple example, if age is a prognostic factor and one study group has more older participants than the other, the allocation scheme is such that the next several older participants would most likely be randomized to the group which currently has fewer older participants. Various methods can be used as the measure of imbalance in prognostic factors. In general, adaptive stratification methods incorporate several prognostic factors in making an "overall assessment" of the group balance or lack of balance. Participants are then assigned to a group in a manner, which will tend to correct an existing imbalance or cause the least imbalance in prognostic factors. This method is sometimes called *minimization* because imbalances in the distribution of prognostic factors are minimized. However, as indicated in the Appendix, the term minimization is also used to refer to a very specific form of adaptive stratification [51, 73]. Generalization of this strategy exists for more than two study groups. Development of these methods was motivated in part by the previously described problems with nonadaptive stratified randomization for small studies. Adaptive methods do not have empty or near empty strata because randomization does not take place within a stratum although prognostic factors are used. Minimization gives unbiased estimates of treatment effect and slightly increased power relative to stratified randomization [73]. These methods are being used, especially in clinical trials of cancer where several prognostic factors need to be balanced, and the sample size is typically 100–200 participants.

The major advantage of this procedure is that it protects against a severe baseline imbalance for important prognostic factors. Overall marginal balance is maintained in the intervention groups with respect to a large number of prognostic factors. One disadvantage is that adaptive stratification is operationally more difficult to carry out, especially if a large number of factors are considered. Although White and Freedman [52] initially developed a simplified version of the adaptive stratification method by using a set of specially arranged index cards, today any small programmable computer can easily carry out the calculations. In addition, the population recruited needs to be stable over time, just as for other adaptive methods. For example, if treatment guidelines change during a long recruitment period, necessitating a change in the inclusion or exclusion criteria, the adaptive procedure may not be able to correct imbalances that developed beforehand. This happened in the Stop Atherosclerosis in Native Diabetics Study (SANDS), a trial comparing intensive intervention for cholesterol and blood pressure with less intensive intervention in people with diabetes [74, 75]. Randomization was done using the urn design, but partway through the trial, new and more aggressive guidelines regarding lipid lowering treatment in people who had known coronary heart disease came out. The participants in SANDS who met those guidelines could no longer be treated with the less intensive regimen and no new participants who had had prior cardiovascular events could be enrolled. Not only was there a possibility of imbalance between

study groups, but the sample size also needed to be reconsidered because of the lower average risk level of the participants. Another disadvantage of adaptive randomization is that the data analysis is complicated, from a strict viewpoint, by the randomization process. The appropriate analysis involves simulating on a computer the assignment of participants to groups by the actual adaptive strategy used. Replication of the simulation, assuming that no group differences exist, generates the significance level of the statistical test to be used.

Biostatisticians are not likely to go through the simulation experiments but would rather use the conventional statistical test and standard critical values to determine significance levels. As with other nonsimple randomization procedures, this strategy is probably somewhat conservative. The impact of one minimization approach on the significance level has been studied [53]. For this case, the authors concluded that if minimization adaptive stratification is used, an analysis of covariance should be employed. To obtain the proper significance level, the analysis should incorporate the same prognostic factors used in the randomization. Minimization and stratification on the same prognostic factors produce similar levels of power, but minimization may add slightly more power if stratification does not include all of the covariates.

Response Adaptive Randomization

Response adaptive randomization uses information on participant response to intervention during the course of the trial to determine the allocation of the next participant. Examples of response adaptive randomization models are the Play the Winner [76] and the two-armed bandits [77] models. These models assume that the investigator is randomizing participants to one of two interventions and that the primary response variable can be determined quickly relative to the total length of the study. Bailar [78] and Simon [79] reviewed the uses of these stratification methods. Additional modifications or methods were developed [80–84].

The *Play the Winner* procedure may assign the first participant by the toss of a coin. The next participant is assigned to the same group as the first participant if the response to the intervention was a success; otherwise, the participant is assigned to the other group. That is, the process calls for staying with the winner until a failure occurs and then switching. The following example illustrates a possible randomization scheme where S indicates intervention success and F indicates failure:

Assignment	Participant								
	1	2	3	4	5	6	7	8	…
Group *A*	S	F				S	F		
Group *B*			S	S	F			S	

Another response adaptive randomization procedure is the *two-armed bandit* method, which continually updates the probability of success as soon as the outcome

for each participant is known. That information is used to adjust the probabilities of being assigned to either group in such a way that a higher proportion of future participants would receive the currently "better" or more successful intervention.

Both of these response adaptive randomization methods have the intended purpose of maximizing the number of participants on the "superior" intervention. They were developed in response to ethical concerns expressed by some clinical investigators about the randomization process. Although these methods do maximize the number of participants on the "superior" intervention, the possible imbalance will almost certainly result in some loss of power and require more participants to be enrolled into the study than would a fixed allocation with equal assignment probability [85]. A major limitation is that many clinical trials do not have an immediately occurring response variable. They also may have several response variables of interest with no single outcome easily identified as being the one upon which randomization should be based. Furthermore, these methods assume that the population from which the participants are drawn is stable over time. If the nature of the study population should change and this is not accounted for in the analysis, the reported significance levels could be biased, perhaps severely [86]. Here, as before, the data analysis should ideally take into account the randomization process employed. For response adaptive methods, that analysis will be more complicated than it would be with simple randomization. Because of these disadvantages, response adaptive procedures are not commonly used.

One application of response adaptive allocation can be found in a trial evaluating extracorporeal membrane oxygenator (ECMO) in a neonatal population suffering from respiratory insufficiency [83, 84, 87–89]. This device oxygenates the blood to compensate for the inability or inefficiency of the lungs to achieve this task. In this trial, the first infant was allocated randomly to control therapy. The result was a failure. The next infant received ECMO, which was successful. The next ten infants were also allocated to ECMO and all outcomes were successful. The trial was then stopped. However, the first infant was much sicker than the ECMO-treated infants. Controversy ensued and the benefits of ECMO remain unclear. This experience does not offer encouragement to use this adaptive randomization methodology.

Mechanics of Randomization

The manner in which the chosen randomization method is actually implemented is very important [90]. If this aspect of randomization does not receive careful attention, the entire randomization process can easily be compromised, thus voiding any of the advantages for using it. To accomplish a valid randomization, it is recommended that an independent central unit be responsible for developing the randomization process and making the assignments of participants to the appropriate group. For a single center trial, this central unit might be a biostatistician or clinician not involved with the care of the participants. In the case of a multicenter trial, the

randomization process is usually handled by the data coordinating center. Ultimately, however, the integrity of the randomization process will rest with the investigator.

Chalmers and colleagues [91] reviewed the randomization process in 102 clinical trials, 57 where the randomization was unknown to the investigator and 45 where it was known. The authors reported that in 14% of the 57 studies, at least one baseline variable was not balanced between the two groups. For the studies with known randomization schedules, twice as many, or 26.7%, had at least one prognostic variable maldistributed. For 43 nonrandomized studies, such imbalances occurred four times as often or in 58%. The authors emphasized that those recruiting and entering participants into a trial should not be aware of the next intervention assignment.

In many cases when a fixed proportion randomization process is used, the randomization schedules are made before the study begins [92–96]. The investigators may call a central location, and the person at that location looks up the assignment for the next participant [92]. Another possibility, used historically and still sometimes in trials involving acutely ill participants, is to have a scheme making available sequenced and sealed envelopes containing the assignments [93]. As a participant enters the trial, she receives the next envelope in the sequence, which gives her the assignment. Envelope systems, however, are more prone to errors and tampering than the former method. In one study, personnel in a clinic opened the envelopes and arranged the assignments to fit their own preferences, accommodating friends and relatives entering the trial. In another case, an envelope fell to the bottom of the box containing the envelopes, thus changing the sequence in which they were opened. Many studies prefer the telephone system to protect against this problem. In an alternative procedure that has been used in several double-blind drug studies, medication bottles are numbered with a small perforated tab [96]. The bottles are distributed to participant in sequence. The tab, which is coded to identify the contents, is torn off and sent to the central unit. This system is also subject to abuse unless an independent person is responsible for dispensing the bottles. Many clinical trials using a fixed proportion randomization schedule require that the investigator call the central location to verify that a participant is eligible to be in the trial before any assignment is made. This increases the likelihood that only eligible participants will be randomized.

For many trials, especially multicenter and multinational trials, logistics require a central randomization operations process. This may be achieved by logging in to a central computer via the internet. In some cases, the clinic may register a participant by dialing into a central computer and entering data via touchtone, with a voice response. These systems, referred to as Interactive Voice Response Systems or IVRS, or Interactive Web Response Systems, IWRS, are effective and can be used to not only assign intervention but can also capture basic eligibility data. Before intervention is assigned, baseline data can be checked to determine eligibility. This concept has been used in a pediatric cancer cooperative clinical trial network [97] and in major multicenter trials [98, 99]. The web-based IWRS systems are becoming common.

Whatever system is chosen to communicate the intervention assignment to the investigator or the clinic, the intervention assignment should be given as closely as possible to the moment when both investigator and participant are ready to begin

the intervention. If the randomization takes place when the participant is first identified and the participant withdraws or dies before the intervention actually begins, a number of participants will be randomized before being actively involved in the study. An example of this occurred in a nonblinded trial of alprenolol in survivors of an acute myocardial infarction [100]. In that trial, 393 participants with a suspected myocardial infarction were randomized into the trial at the time of their admission to the coronary care unit. The alprenolol or placebo was not initiated until 2 weeks later. Afterwards, 231 of the randomized participants were excluded because a myocardial infarction could not be documented, death had occurred before therapy was begun, or various contraindications to therapy were noted. Of the 162 participants who remained, 69 were in the alprenolol group and 93 were in the placebo group. This imbalance raised concerns over the comparability of the two groups and possible bias in reasons for participant exclusion. By delaying the randomization until initiation of therapy, the problem of these withdrawals could have been avoided.

Recommendations

For large studies involving more than several hundred participants, the randomization should be blocked. If a large multicenter trial is being conducted, randomization should be stratified by center. Randomization stratified on the basis of other factors in large studies is usually not necessary, because randomization tends to make the study groups quite comparable for all risk factors. The participants can still, of course, be stratified once the data have been collected and the study can be analyzed accordingly.

For small studies, the randomization should also be blocked and stratified by center if more than one center is involved. Since the sample size is small, a few strata for important risk factors may be defined to assure that balance will be achieved for at least those factors. For a larger number of prognostic factors, the adaptive stratification techniques should be considered and the appropriate analyses performed. As in large studies, stratified analysis can be performed even if stratified randomization was not done. For many situations, this will be satisfactory.

Appendix

Adaptive Randomization Algorithm

Adaptive randomization can be used for more than two intervention groups, but for the sake of simplicity only two will be used here. To describe this procedure in more detail, a minimum amount of notation needs to be defined. First, let

x_{ik} = the number of participants already assigned intervention k ($k = 1,2$) who have the same level of prognostic factor i ($i = 1,2,\ldots, f$) as the new participant

And define

$$x_{ik}^{t} = x_{ik} \quad \text{if } t \neq k$$

$$= x_{ik} + 1 \quad \text{if } t = k$$

The x_{ik}^{t} represents the change in balance of allocation if the new participant is assigned intervention *t*. Finally, let

$B(t)$ = function of the x_{ik}^{t}s, which measures the "lack of balance" over all prognosticfactors if the next participant is assigned intervention *t*.

Many possible definitions of $B(t)$ can be identified. As an illustrative example, let

$$B(t) = \sum_{i=1}^{f} w_i \,\text{Range}\,(x_{i1}^{t}, x_{i2}^{t})$$

where w_i = the relative importance of factor i to the other factors and the range is the absolute difference between the largest and smallest values of x_{i1}^{t} *and* x_{i2}^{t}.

The value of $B(t)$ is determined for each intervention ($t = 1$ and $t = 2$). The intervention with the smaller $B(t)$ is preferred, because allocation of the participant to that intervention will cause the least imbalance. The participant is assigned, with probability $p > 1/2$, to the intervention with the smaller score, $B(1)$ or $B(2)$. The participant is assigned, with probability $(1-p)$, to the intervention with the larger score. These probabilities introduce the random component into the allocation scheme. Note that if $p = 1$ and, therefore, $1-p = 0$, the allocation procedure is deterministic (no chance or random aspect) and has been referred to by the term "minimization" [51, 53].

As a simple example of the adaptive stratification method, suppose there are two groups and two prognostic factors to control. The first factor has two levels and the second factor has three levels. Assume that 50 participants have already been randomized and the following table summarizes the results (Table 6.2).

Table 6.2 Fifty randomized participants by group and level of factor (*xik*s)[a]

Factor	1		2			
Level	1	2	1	2	3	
Group						Total
1	16	10	13	9	4	26
2	14	10	12	6	6	24
	30	20	25	15	10	50

[a]After Pocock and Simon [34]

In addition, the function $B(t)$ as defined above will be used with the range of the x^1_{ik} s as the measure of imbalance, where $w_1=3$ and $w_2=2$; that is, the first factor is 1.5 times as important as the second as a prognostic factor. Finally, suppose $p=2/3$ and $1-p=1/3$.

If the next participant to be randomized has the first level of the first factor and the third level of the second factor, then this corresponds to the first and fifth columns in the table. The task is to determine $B(1)$ and $B(2)$ for this participant as shown below.

(a) Determine $B(1)$

	Factor 1, Level 1			
	K	x_{1k}	x^1_{1k}	Range (x^1_{11}, x^1_{12})
Group	1	16	17	$\|17-14\|=3$
	2	14	14	
	Factor 2, Level 3			
	K	x_{2k}	x^1_{2k}	Range (x^1_{21}, x^1_{22})
Group	1	4	5	$\|5-6\|=1$
	2	6	6	

Using the formula given, *B(1)* is computed as $3\times3+2\times1=11$.

(b) Determine *B(2)*

	Factor 1, Level 1			
	K	x_{1k}	x^2_{1k}	Range (x^2_{11}, x^2_{12})
Group	1	16	16	$\|16-15\|=1$
	2	14	15	
	Factor 2, Level 3			
	K	x_{2k}	x^2_{1k}	Range (x^2_{21}, x^2_{22})
Group	1	4	4	$\|4-7\|=3$
	2	6	7	

Then *B(2)* is computed as $3\times1+2\times3=9$.

(c) Now rank *B(1)* and *B(2)* from smaller to larger and assign with probability p the group with the smaller $B(t)$.

t	$B(t)$	Probability of assigning t
2	$B(2)=9$	$p=2/3$
1	$B(1)=11$	$1-p=1/3$

Thus, this participant is randomized to Group 2 with probability 2/3 and to Group 1 with probability 1/3. Note that if minimization were used ($p=1$), the assignment would be Group 2.

References

1. Hill AB. The clinical trial. *Br Med Bull* 1951;7:278–282.
2. Armitage P. The role of randomization in clinical trials. *Stat Med* 1982;1:345–352.

3. Byar DP, Simon RM, Friedewald WT, et al. Randomized clinical trials: Perspectives on some recent ideas. *N Engl J Med* 1976;295:74–80.
4. Zelen M. The randomization and stratification of patients to clinical trials. *J Chronic Dis* 1974;27:365–375.
5. Pocock SJ. Allocation of patients to treatment in clinical trials. *Biometrics* 1979;35:183–197.
6. Peto R. Clinical trial methodology. *Biomedicine* 1978;28(special issue):24–36.
7. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomised clinical trials requiring prolonged observation of each patient. 1. Introduction and design. *Br J Cancer* 1976;34: 585–612.
8. Brown BW. Statistical controversies in the design of clinical trials – some personal views. *Control Clin Trials* 1980;1:13–27.
9. Lachin JM. Statistical properties of randomization in clinical trials. *Control Clin Trials* 1988;9:289–311.
10. Lachin JM, Matts JP, Wei LJ. Randomization in clinical trials: Conclusions and recommendations. *Control Clin Trials* 1988;9:365–374.
11. Royall RM. Ethics and statistics in randomized clinical trials. *Stat Sci* 1991;6(1):52–88.
12. Weinstein MC. Allocation of subjects in medical experiments. *N Engl J Med* 1974;291: 1278–1285.
13. Bather JA. On the allocation of treatments in sequential medical trials. *Int Stat Rev* 1985;53:1–13.
14. Kalish LA, Begg CB. Treatment allocation methods in clinical trials: A review. *Stat Med* 1985;4:129–144.
15. Stigler SM. The use of random allocation for the control of selection bias. *Biometrika* 1969;56:553–560.
16. Wei LJ. On the random allocation design for the control of selection bias in sequential experiments. *Biometrika* 1978;65:79–84.
17. Altman D, Dore CJ. Randomization and baseline comparisons in clinical trials. *Lancet* 1990;335:149–155.
18. Williams DS, Davis CE. Reporting of assignment methods in clinical trials. *Control Clin Trials* 1994;15:294–298.
19. Moher D, Schulz KF, Altman DG, CONSORT Group (Consolidated Standards of Reporting Trials). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Ann Intern Med* 2001;134:657-662.
20. Mills EJ, Wu P, Gagnier J, Devereaux PJ. The quality of randomized trial reporting in leading medical journals since the revised CONSORT statement. *Contemp Clin Trials* 2005;26: 480–487.
21. Brittain E, Schlesselman JJ. Optimal allocation for the comparison of proportions. *Biometrics* 1982;38:1003–1009.
22. Lachin JM. Properties of simple randomization in clinical trials. *Control Clin Trials* 1988;9:312–326.
23. Louis TA. Optimal allocation in sequential tests comparing the means of two Gaussian populations. *Biometrika* 1975;62:359–369.
24. Louis TA. Sequential allocation in clinical trials comparing two exponential survival curves. *Biometrics* 1977;33:627–634.
25. Kalish LA, Harrington DP. Efficiency of balanced treatment allocation for survival analysis. *Biometrics* 1988;44:815–821.
26. Matts JP, Lachin JM. Properties of permutated-block randomization in clinical trials. *Control Clin Trials* 1988;9:327–344.
27. Kalish LA, Begg CB. The impact of treatment allocation procedures on nominal significance levels and bias. *Control Clin Trials* 1987;8:121–135.
28. Smythe RT, Wei LJ. Significance tests with restricted randomization design. *Biometrika* 1983;70:496–500.
29. Steele JM. Efron's conjecture on vulnerability to bias in a method for balancing sequential trials. *Biometrika* 1980;67:503–504.

30. Titterington DM. On constrained balance randomization for clinical trials. *Biometrics* 1983;39:1083–1086.
31. Matts JP, McHugh RB. Analysis of accrual randomized clinical trials with balanced groups in strata. *J Chronic Dis* 1978;31:725–740.
32. Zelen M. Aspects of the planning and analysis of clinical trials in cancer. In Srivastava JN (ed.). *A Survey of Statistical Design and Linear Models.* Amsterdam: North-Holland, 1975.
33. Coronary Drug Project Research Group. Factors influencing long term prognosis after recovery from myocardial infarction – Three year findings of the Coronary Drug Project. *J Chronic Dis* 1974;27:267–285.
34. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 1975;31:103–115.
35. Green SB, Byar DP. The effect of stratified randomization on size and power of statistical tests in clinical trials. *J Chronic Dis* 1978;31:445–454.
36. Ducimetiere P. Stratification. In Boissel JP, Klimt CR (eds.). *Multi-center Controlled Trials: Principals and Problems.* Paris: INSERM, 1979.
37. Simon R. Restricted randomization designs in clinical trials. *Biometrics* 1979;35:503–512.
38. Meier P. Stratification in the design of a clinical trial. *Control Clin Trials* 1981;1:355–361.
39. Grizzle JE. A note on stratifying versus complete random assignment in clinical trials. *Control Clin Trials* 1982;3:365–368.
40. McHugh R, Matts J. Post-stratification in the randomized clinical trial. *Biometrics* 1983;39:217–225.
41. Fleiss JL. Multicentre clinical trials: Bradford Hill's contributions and some subsequent developments. *Stat Med* 1982;1:353–359.
42. Feinstein AR, Landis JR. The role of prognostic stratification in preventing the bias permitted by random allocation of treatment. *J Chronic Dis* 1976;29:277–284.
43. Mantel N. Pre-stratification or post-stratification (Letter). *Biometrics* 1984;40:256–258.
44. Palta M. Investigating maximum power losses in survival studies with nonstratified randomization. *Biometrics* 1985;41:497–504.
45. Palta M, Amini SB. Magnitude and likelihood of loss resulting from non-stratified randomization. *Stat Med* 1982;1:267–275.
46. Aspirin Myocardial Infarction Study Research Group. A randomized controlled trial of aspirin in persons recovered from myocardial infarction. *JAMA* 1980;243:661–669.
47. Efron B. Forcing a sequential experiment to be balanced. *Biometrika* 1971;58:403–417.
48. Freedman LS, White SJ. On the use of Pocock and Simon's method for balancing treatment numbers over prognostic factors in the controlled clinical trial. *Biometrics* 1976;32:691–694.
49. Begg CD, Iglewicz B. A treatment allocation procedure for sequential clinical trials. *Biometrics* 1980;36:81–90.
50. Atkinson AC. Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika* 1982;69:61–67.
51. Taves DR. Minimization: A new method of assigning patients to treatment and control groups. *Clin Pharmacol Ther* 1974;15:443–453.
52. White SJ, Freedman LS. Allocation of patients to treatment groups in a controlled clinical study. *Br J Cancer* 1978;37:849–857.
53. Forsythe AB, Stitt FW. Randomization or minimization in the treatment assignment of patient trials: validity and power of tests. Technical Report No. 28, Health Science Computer Facility, University of California, Los Angeles, 1977.
54. Begg CB. On inferences from Wei's biased coin design for clinical trials. *Biometrika* 1990;77:467–484.
55. Efron B. Randomizing and balancing a complicated sequential experiment. In Miller RG Jr. Efron B, Brown BW Jr, Moses LE (eds.). *Biometrics Casebook.* New York: Wiley, 1980, pp. 19–30.
56. Halpern J, Brown BW Jr. Sequential treatment allocation procedures in clinical trials – with particular attention to the analysis of results for the biased coin design. *Stat Med* 1986;5: 211–229.

57. Hannigan JR Jr, Brown BW Jr. Adaptive randomization based coin-design: Experience in a cooperative group clinical trial. Technical Report 74, Division of Biostatistics, Stanford University, Stanford, California, 1982.
58. Klotz JH. Maximum entropy constrained balance randomization for clinical trials. *Biometrics* 1978;34:283–287.
59. Raghavaro D. Use of distance function in sequential treatment assignment for prognostic factors in the controlled clinical trial. *Calcutta Stat Assoc Bull* 1980;29:99–102.
60. Smith RL. Sequential treatment allocation using biased coin designs. *J R Stat Soc Series B Stat Methodol* 1984;46:519–543.
61. Soares JF, Wu CFJ. Some restricted randomization rules in sequential designs. *Commun Stat Theory Methods A* 1983;12:2017–2034.
62. Wei LJ. The adaptive biased coin design for sequential experiments. *Ann Stat* 1978;6: 92–100.
63. Wei LJ. A class of designs for sequential clinical trials. *J Am Stat Assoc* 1977;72:382–386.
64. Wei LJ. A class of treatment assignment rules for sequential experiments. *Commun Stat Theory Methods A* 1978;7:285–295.
65. Wei LJ, Lachin JM. Properties of the urn randomization in clinical trials. *Control Clin Trials* 1988;9:345–364.
66. Wei LJ, Smythe RT, Lin DY, Park TS. Statistical inferences with data-dependent treatment allocation rules. *J Am Stat Assoc* 1990;85:156–162.
67. Wei LJ, Smythe RT, Smith RL. K-treatment comparisons with restricted randomization rules in clinical trials. *Ann Stat* 1986;14:265–274.
68. Wei LJ. An application of an urn model to the design of sequential controlled clinical trials. *J Am Stat Assoc* 1978;73:559–563.
69. The DCCT Research Group. Diabetes Control and Complications Trial (DCCT): Design and methodologic considerations for the feasibility phase. *Diabetes* 1986;35:530–545.
70. Begg CB, Kalish LA. Treatment allocation for nonlinear models in clinical trials: The logistic model. *Biometrics* 1984;40:409–420.
71. Begg CB, Kalish LA. Treatment allocation in sequential clinical trials: Nonlinear models. *Proc Stat Comput Sect, Am Stat Assoc* 1982:57–60.
72. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984;71:431–444.
73. Birkett JJ. Adaptive allocation in randomized controlled trials. *Control Clin Trials* 1985;6:146–155.
74. Russell M, Fleg JL, Galloway J, et al. Examination of lower targets for low-intensity lipoprotein cholesterol and blood pressure in diabetes—the Stop Atherosclerosis in Native Diabetics Study (SANDS). *Am Heart J* 2006:152;867–875.
75. Howard BV, Roman MJ, Devereux RB, et al. Effect of lower targets for blood pressure and LDL cholesterol on atherosclerosis in diabetes: The SANDS randomized trial. *JAMA* 2008:299;1678–1689.
76. Zelen M. Play-the-winner rule and the controlled clinical trial. *J Am Stat Assoc* 1969;64: 131–146.
77. Robbins H. Some aspects of the sequential design of experiments. *Bull Am Math Soc* 1952;58:527–535.
78. Bailar JC. Patient assignment algorithms: An overview. In *Proceedings of the 9th International Biometric Conference,* Raleigh, NC: The Biometric Society, 1976; Vol I, pp. 189–206.
79. Simon R. Adaptive treatment assignment methods and clinical trials. *Biometrics* 1977;33: 743–749.
80. Armitage P. The search for optimality in clinical trials. *Int Stat Rev* 1985;53:15–24.
81. Nordbrock E. An improved play-the-winner sampling procedure for selecting the better of two binomial populations. *J Am Stat Assoc* 1976;71:137–139.
82. Wei LJ. Exact two-sample permutation tests based on the randomized play-the-winner rule. *Biometrika* 1988;75:603–606.

83. Bartlett RH, Roloff DW, Cornell RG, et al. Extracorporeal circulation in neonatal respiratory failure: A prospective randomized study. *Pediatrics* 1985;76:479–487.
84. O'Rourke PP, Crone RK, Vacanti JP, et al. Extracorporeal membrane oxygenation and conventional medical therapy in neonates with persistent pulmonary hypertension of the newborn: A prospective randomized study. *Pediatrics* 1989;84:957–963.
85. Simon R, Weiss GH, Hoel DG. Sequential analysis of binomial clinical trials. *Biometrika* 1975;62:195–200.
86. Simon R, Hoel DG, Weiss GH. The use of covariate information in the sequential analysis of dichotomous response experiments. *Commun Stat Theory Methods* 1977;8:777–788.
87. Paneth N, Wallenstein S. Extracorporeal membrane oxygenation and the play the winner rule. *Pediatrics* 1985;76:622–623.
88. Ware JH. Investigating therapies of potentially great benefit: ECMO. *Stat Sci* 1989;4:298–340.
89. Ware JH, Epstein MF. Extracorporeal circulation in neonatal respiratory failure: A prospective randomized study. *Pediatrics* 1985;76:849–851.
90. Pocock SJ, Lagakos SW. Practical experience of randomization in cancer trials: An international survey. *Br J Cancer* 1982;46:368–375.
91. Chalmers TC, Celano P, Sacks HS, et al. Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 1983;309:1358–1361.
92. Beta-Blocker Heart Attack Trial Research Group. A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. *JAMA* 1982;247:1707–1714.
93. Hypertension Detection and Follow-up Program Cooperative Group. Five-year findings of the Hypertension Detection and Follow-up Program. Reduction in mortality of persons with high blood pressure, including mild hypertension. *JAMA* 1979;242:2562–2571.
94. Multiple Risk Factor Intervention Trial Research Group. Multiple Risk Factor Interventional Trial. Risk factor changes and mortality results. *JAMA* 1982;248:1465–1477.
95. CASS Principal Investigators and Their Associates. Coronary Artery Surgery Study (CASS): A randomized trial of coronary artery bypass surgery, survival data. *Circulation* 1983;68: 939–950.
96. Collaborative Group on Antenatal Steroid Therapy. Effect of antenatal dexamethasone administration on the prevention of respiratory distress syndrome. *Am J Obstet Gynecol* 1981;141:276–287.
97. Krischer J, Hurley C, Pillamarri M, et al. An automated patient registration and treatment randomization system for multicenter clinical trials. *Control Clin Trials* 1991;12:367–377.
98. Kjekshus J, Apetrei E, Barrios V, et al. Rosuvastatin in older patients with systolic heart failure. *N Engl J Med* 2007;357:2248–2261.
99. SPORTIF Executive Steering Committee for the SPORTIF-V Investigators. Ximelagatran vs Warfarin for stroke prevention in patients with nonvalvular atrial fibrillation. A randomized trial. *JAMA* 2005;293:690–698.
100. Ahlmark G, Saetre H. Long-term treatment with β-blockers after myocardial infarction. *Eur J Clin Pharmacol* 1976;10:77–83.

Chapter 7
Blindness

In any clinical trial, bias is one of the main concerns. Bias may be defined as systematic error, or "difference between the true value and that actually obtained due to all causes other than sampling variability" [1]. It can be caused by conscious factors, subconscious factors, or both. Bias can occur at a number of places in a clinical trial, from the initial design through data analysis and interpretation. One general solution to the problem of bias is to keep the participant and the investigator blinded, or masked, to the identity of the assigned intervention. One can also blind several other aspects of a trial including the assessment, classification, and evaluation of the response variables. Large sample size does not reduce bias although it generally improves precision and thus power.

The blinding terminology is not well understood. A survey of 91 internal medicine physicians in Canada [2] showed that 75% knew the definition of single-blind. Approximately 40% understood the proper definition of double-blind and less than 20% could define triple-blind. The use of the terms single-blind and double-blind is particularly inconsistent in trials of non-pharmaceutical interventions [3].

Fundamental Point

A clinical trial should, ideally, have a double-blind design in order to avoid potential problems of bias during data collection and assessment. In studies where such a design is impossible, other measures to reduce potential bias are advocated.

Types of Trials

Unblinded

In an unblinded or open trial, both the participant and the investigator know to which intervention the participant has been assigned. Some kinds of trials can be conducted only in this manner and include those involving most surgical procedures,

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_7, © Springer Science+Business Media, LLC 2010

comparisons of devices and medical treatment, changes in lifestyle (e.g., eating habits, exercise, cigarette smoking), or learning techniques.

An unblinded study is appealing for two reasons. First, all other things being equal, it is simpler to execute than other studies. The usual drug trial may be easier to design and carry out, and consequently less expensive if blinding is not an issue. Also, it has been argued that it more accurately reflects clinical practice [4]. However, an unblinded trial need not be simple – for example, trials that simultaneously attempt to induce lifestyle changes and test drug interventions, such as the Women's Health Initiative (WHI) [5], one of the three interventions of which was an unblinded dietary intervention. It involved three distinct interventions: a low-fat eating pattern, hormone replacement therapy, and calcium and vitamin D supplementation. Second, investigators are likely to be more comfortable making decisions, such as whether or not to continue a participant on his assigned study medication if they know its identity.

The main disadvantage of an unblinded trial is the possibility of bias. Participant reporting of symptoms and side effects and prescription of concomitant or compensatory treatment are all susceptible to bias. Other problems of biased data collection and assessment by the investigator are addressed in Chap. 11. Since participants when joining a trial have sincere hopes and expectations about beneficial effects, they may become dissatisfied and drop-out of the trial in disproportionately large numbers if not on the new or experimental intervention.

A trial of the possible benefits of ascorbic acid (vitamin C) in the common cold was designed as a double-blind study [6, 7]. However, it soon became apparent that many of the participants, most of whom were medical staff, discovered mainly by tasting whether they were on ascorbic acid or placebo. As more participants became aware of their medication's identity, the dropout rate in the placebo group increased. Since evaluation of severity and duration of colds depended on the participant's reporting of his or her symptoms, this unblinding was important. Among those participants who claimed not to know the identity of the treatment, ascorbic acid showed no benefit over placebo. In contrast, among participants who knew or guessed what they were on, ascorbic acid did better than placebo. Therefore, preconceived notions about the benefit of a treatment, coupled with a subjective response variable, may have yielded biased reporting. The investigators' willingness to share this experience provided us with a nice illustration of the importance of maintaining blindness.

In a trial of coronary artery bypass surgery versus medical treatment [8], the number of participants who smoked was equal in the two study groups at baseline. During the early part of follow-up, there were significantly fewer smokers in the surgical group than in the medical group. The effect of this group difference on the outcome of the trial is difficult, if not impossible, to assess.

Single-Blind

In a single-blind study, only the investigator is aware of which intervention each participant is receiving. The advantages of this design are similar to those of an unblinded

study – it is usually simpler to carry out than a double-blind design, and knowledge of the intervention may help the investigator exercise her best judgment when caring for the participants. Indeed, certain investigators are reluctant to participate in studies in which they do not know the study group assignment. They may recognize that bias is partially reduced by keeping the participant blinded but feel that the participant's health and safety are best served if they themselves are not blinded.

The disadvantages of a single-blind design are similar to, though not so pronounced as, those of an unblinded design. The investigator avoids the problems of biased participant reporting, but she herself can affect the administration of nonstudy therapy, data collection, and data assessment. For example, a single-blind study reported benefits from zinc administration in a group of people with taste disorders [9]. Because of the possibility of bias in a study using a response variable as subjective and hard to measure as taste, the study was repeated, using a type of crossover, double-blind design [10]. This second study showed that zinc, when compared with placebo, did not relieve the taste disorders of the study group. The extent of the blinding of the participants did not change; therefore, presumably, knowledge of drug identity by the investigator was important. The results of treatment cross-over were equally revealing. In the single-blind study, participants who did not improve when given placebo as the first treatment, "improved" when placed on zinc. However, in all four double-blind, cross-over procedures (placebo to zinc, placebo to placebo, zinc to zinc, zinc to placebo), the participants who had previously shown no improvement on the first treatment did show benefit when given the second medication. Thus, the expectation that the participants who failed to respond to the first drug were now being given an active drug may have been sufficient to produce a positive response.

A more recent example comes from two noninferiority trials comparing ximelagatran, a novel oral direct thrombin inhibitor, to warfarin for the prevention of thromboembolic events in people with nonvalvular atrial fibrillation [11]. The first trial, SPORTIF III, was single-blind with blinded events assessment, while the second trial, SPORTIF V, was double-blind. The primary response variable was all strokes and systemic embolic events. The observed risk ratio in the single-blind SPORTIF III was 0.71 (95% CI, 0.48–1.07) while the result trended in the opposite direction in the double-blind SPORTIF V with a risk ratio of 1.38 (95% CI, 0.91–2.10). One cannot be sure how much bias may have played a role, but, in general, more confidence ought to be placed on trials with a double-blind design.

Both unblinded and single-blind trials are vulnerable to another source of potential bias introduced by the investigators. This relates to group differences in compensatory and concomitant treatment. Investigators may feel that the control group is not being given the same opportunity as the intervention group and, as a result, may prescribe additional treatment as "compensation." This may be in the form of advice or therapy. For example, several studies have attempted blood pressure lowering as either the sole intervention or part of a broader effort. In general, the investigators would make an intensive effort to persuade participants in the intervention group to take their study medication. To persuade successfully, the investigators themselves had to be convinced that blood pressure reduction was likely beneficial. When they

were seeing participants who had been assigned to the control group, this conviction was difficult to suppress. Therefore, participants in the control group were likely to have been instructed about non-pharmacological ways by which to lower their blood pressure. The result of compensatory treatment is a diminution of the difference between the intervention group and the "untreated," or control group.

Working against this is the fact that investigators prefer to be associated with a study that gives positive findings. Favorable results published in a reputable journal are likely to lead to more invitations to present the findings at scientific meetings and grand rounds and can also support academic promotions. Investigators may, therefore, subconsciously favor the intervention group when they deal with participants, collect data, and assess results.

Concomitant treatment means any non-study therapy administered to participants during a trial. If such treatment is likely to influence the response variable, this needs to be considered when determining sample size. Of more concern is the bias that can be introduced if concomitant treatment is applied unequally in the two groups. In order to bias the outcome of a trial, concomitant treatment must be effective, and it must be used in a high proportion of the participants. When this is the case, bias is a possibility and may occur in either direction, depending on whether the concomitant treatment is preferentially used in the control, or in the intervention group. It is usually impossible to determine the direction and magnitude of such bias in advance or its impact after it has occurred.

Double-Blind

In a double-blind study, neither the participants nor the investigators responsible for following the participants, collecting data, and assessing outcomes should know the identity of the intervention assignment. Such designs are usually restricted to trials of drug or biologics. It is theoretically possible to design a study comparing two surgical procedures or implantation of two devices in which the surgeon performing the operation knows the type of surgery or device, but neither the study investigator nor the participant knows. Similarly, one might be able to design a study comparing two diets in which the food looks identical. However, such trials are uncommon.

The main advantage of a truly double-blind study is that the risk of bias is reduced. Preconceived ideas of the investigator will be less important because he or she will not know which intervention a particular participant receives. Any effect of her actions, therefore, would theoretically occur equally in the intervention and control groups. As discussed later, the possibility of bias may never be completely eliminated. However, a well designed and properly run double-blind study can minimize bias. As in the example of the trial of zinc and taste impairment, double-blind studies have at times led to results that differ from unblinded or single blind studies. Such cases illustrate the role of bias as a factor in clinical trials.

The double-blind design is no protection against imbalances in use of concomitant medications. A placebo-controlled trial of a long-acting inhaled anticholinergic

medication in participants with chronic obstructive pulmonary disease allowed the use of any other available drug treatment for this condition as well as a short-acting inhaled anticholinergic agent for acute exacerbations [12]. The extent of this co-intervention is likely to differ between the actively treated and the placebo groups, but the findings by study group were not presented. Moreover, it may influence symptomology as well as risks of disease events and make it very difficult to determine the true effects of the long-acting anticholinergic inhaler. Reporting the proportion of participants given a co-intervention at any time over the 4 years of the trial by treatment group would help the interpretation of results, but this does not take into account the frequency and intensity of its use.

In a double-blind trial, certain functions, which in open or single-blind studies could be accomplished by the investigators, must be taken over by others in order to maintain the blindness. Thus, an outside body needs to monitor the data for toxicity and benefit, especially in long-term trials. Chapter 16 discusses data monitoring in greater detail. A person other than the investigator who sees the participants needs to be responsible for assigning the interventions to the participants and monitoring them for safety. Treatments that require continuous dose adjustment, such as warfarin, are difficult to blind, but it can be accomplished. In one trial [13], an unblinded pharmacist or physician adjusted not only the warfarin doses according to an algorithm for maintaining the International Normalized Ratio (INR), a measure of anticoagulation, within a pre-specified range, but also the placebo doses randomly. The authors concluded that "placebo warfarin dose adjustment schedules can protect blinding adequately" for participants and investigators and recommended their use for future trials of warfarin. A similar approach was employed in the Coumadin Aspirin Reinfarction Study [14]. An INR control center adjusted the doses in the three treatment arms to keep the INR values below the prespecified safety limits and to maintain the double-blind.

In many single- and double-blind drug trials, the control group is placed on a matched placebo. Much debate has centered on the ethics of using a placebo. See Chap. 2 for a further discussion of this issue.

Triple-Blind

A triple-blind study is an extension of the double-blind design; the committee monitoring response variables is not told the identity of the groups. The committee is simply given data for groups A and B. A triple-blind study has the theoretical advantage of allowing the monitoring committee to evaluate the response variable results more objectively. This assumes that appraisal of efficacy and harm, as well as requests for special analyses, may be biased if group identity is known. However, in a trial where the monitoring committee has an ethical responsibility to ensure participant safety, such a design may be counterproductive. When hampered in the safety-monitoring role, the committee cannot carry out its responsibility to minimize harm to the participants, since monitoring is often guided by the constellation of

trends and their directions. In addition, even if the committee could discharge its duties adequately while being kept blinded, many investigators would be uneasy participating in such a study. Though in most cases the monitoring committee looks only at group data and can rarely make informed judgments about individuals, the investigator still relies on the committee to safeguard her study participants. This may not be a completely rational approach because, by the time many monitoring committees receive data, often any emergency situation has long passed. Nevertheless, the discomfort many investigators feel about participating in double-blind studies would be magnified should the data monitoring committee also be kept blinded.

Finally, people tend not to accept beneficial outcomes unless a statistically significant difference has been achieved. Rarely, though, will investigators want to continue a study in order to achieve a clearly significant difference in an adverse direction; that is, until the intervention is statistically significantly worse or more harmful than the control. Therefore, many monitoring committees demand to know which study groups are on which intervention. We agree with the arguments against triple-blind summarized by Meinert [15].

A triple-blind study can be conducted ethically if the monitoring committee asks itself at each meeting whether the direction of observed trends matters. If it does not matter, then the triple-blind can be maintained, at least for the time being. This implies that the monitoring committee can ask to be unblinded at any time it chooses. In the Randomized Aldactone Evaluation Study (RALES), the Data and Safety Monitoring Board was split and several members argued against being blinded [16]. However, triple-blind was employed initially. For most variables, the treatment groups were labeled A and B. Since increased rates of gynecomastia and hyperkalemia would unmask the A and B assignments, these adverse events were labeled X and Y.

Protecting the Double-Blind Design

Double-blind studies are usually more complex and therefore more difficult to carry out than other trials. One must ensure that investigators remain blinded and that any data which conceivably might endanger blindness be kept from them during the study. An effective data monitoring scheme must be set up, and emergency unblinding procedures must be established. These requirements pose their own problems and can increase the cost of a study. In the Aspirin-Myocardial Infarction Study [17], a double-blind trial of aspirin in people with coronary heart disease, the investigators wished to monitor the action of aspirin on platelets. A postulated beneficial effect of aspirin relates to its ability to reduce the aggregation of platelets. Therefore, measuring platelet aggregation provided both an estimate of whether the aspirin treated group was getting a sufficient dose and a basis for measurement of participant adherence. However, tests of platelet aggregation need to be performed shortly after the blood sample is drawn. The usual method is to have a laboratory technician insert the specimen in an aggregometer, add a material such as epinephrine (which, in the absence of aspirin, causes platelets to aggregate), and analyze a curve that is

printed on a paper strip. In order to maintain the blind, the study needed to find a way to keep the technician from seeing the curve. Therefore, a cassette tape-recorder was substituted for the usual paper strip recorder and the indicator needle was covered. These changes required a modification of the aggregometer. All of the 30 clinics required this equipment, so the adjustment was expensive. However, it helped ensure the maintenance of the blind.

Naturally, participants want to be on the "better" intervention. In a drug trial, the "better" intervention usually is presumed to be the new one; in the case of a placebo-control trial, it is presumed to be the active medication. Investigators may also be curious about a drug's identity. For these reasons, consciously or unconsciously, both participants and investigators may try to unblind the medication. Unblinding can be done deliberately by going so far as to have the drug analyzed, or in a less purposeful manner by "accidentally" breaking open capsules, holding pills up to the light, carefully testing them, or by taking any of numerous other actions. In the first case, which may have occurred in the vitamin C study discussed earlier, little can be done to ensure blinding absolutely. Curious participants and investigators can discover many ways to unblind the trial, whatever precautions are taken. Probably, however, the less purposeful unblinding is more common.

Drug studies, in particular, lend themselves to double-blind designs. One of the surest ways to unblind a drug study is to have dissimilar appearing medications. When the treatment identity of one participant becomes known to the investigator, the whole trial is unblinded. Thus, matching of drugs is essential.

Matching of Drugs

Proper matching has received little attention in the literature. A notable exception is the vitamin C study [6, 7] in which of the double-blind was not maintained throughout the trial. One possible reason given by the investigators was that, in the rush to begin the study, the contents of the capsules were not carefully prepared. The lactose placebo could easily be distinguished from ascorbic acid by taste, as the study participants quickly discovered. An early report is equally disturbing [18]. The authors noted that, of 22 studies surveyed, only five had excellent matching between the drugs being tested. A number of features of matching must be considered. A review of 191 randomized placebo-controlled trials from leading general medicine and psychiatry journals showed that 81 (42%) trials reported on the matching of drug characteristics [19]. Only 19 (10%) commented on more than one of the matching features and appearance was, by far, the most commonly reported characteristic. Thus, most reports of drug studies do not indicate how closely tablets or capsules resembled one another, or how great a problem was caused by imperfect matching.

Cross-over studies, where each subject sees both medications, require the most care in matching. Visual discrepancies can occur in size, shape, color, and texture. Ensuring that these characteristics are identical may not be simple. In the case of tablets, dyes or coatings may adhere differently to the active ingredient than to the

placebo, causing slight differences in color or sheen. Agents can also differ in odor. The taste and the local action on the tongue of the active medication are likely to be different than those of the placebo. For example, propranolol is a topical anesthetic which causes lingular numbness if held in the mouth. Farr and Gwaltney reported on problems in matching zinc lozenges against placebo [20]. Because zinc lozenges are difficult to blind, the authors questioned whether studies using zinc for common cold prevention were truly valid. They conducted trials illustrating that if a placebo is inadequately matched, the "unpleasant side effects of zinc" may reduce the perception of cold symptoms.

Drug preparations should be pretested if it is possible. One method is to have a panel of observers unconnected with the study compare samples of the medications. Perfect matches are almost impossible to obtain and some differences are to be expected. However, beyond detecting differences, it is important to assess whether the observers can actually identify the agents. If not, slightly imperfect matches may be tolerated. The investigator must remember that, except in crossover studies, the participant has only one drug and is therefore not able to make a comparison. On the other hand, participants may meet and talk in waiting rooms, or in some other way compare notes or pills. Of course, staff always have the opportunity to compare different preparations and undermine the integrity of a study.

Differences may become evident only after some time due to degradation of the active ingredient. Freshly prepared aspirin is relatively odor free, but after a while, tell-tale acetic acid accumulates. Ginkgo biloba has a distinct odor and a bitter taste. In one trial of Ginkgo, the investigators used coated tablets to mask both odor and taste [21]. The tablets were placed in blister packs to reduce the risk of odor. Quinine was added to the placebo tablets to make them as bitter as the active drug. This approach prevented any known blind-breaking.

The use of substances to mask characteristic taste, color, or odor, as was done in the ginkgo biloba trial mentioned above, is often advocated. Adding vanilla to the outside of tablets may mask an odor; adding dyes will mask dissimilar colors. A substance such as quinine or quassin will impart a bitter taste to the preparations. Not only will these chemical substances mask differences in taste, but they will also effectively discourage participants from biting into a preparation more than once. However, the possibility that they may have toxic effects after long-term use or even cause allergic reactions in a small percent of the participants must be considered. It is usually prudent to avoid using extra substances unless absolutely essential to prevent unblinding of the study.

Less obviously, the weight or specific gravity of the tablets may differ. Matching the agents on all of these characteristics may be impossible. However, if a great deal of effort and money are being spent on the trial, a real attempt to ensure matching makes sense. The investigator also needs to make sure that the containers are identical. Bottles and vials need to be free of any marks other than codes which are indecipherable except with the key.

Sometimes, two or more active drugs are being compared. The ideal method of blinding is to have the active agents look alike, either by formulating them appropriately or possibly by enclosing them in identical capsules. The former may not be possible, and the latter may be expensive or require capsules too large to be practical.

In addition, enclosing tablets in capsules may change the rate of absorption and the time to treatment response. In a comparative acute migraine trial, one manufacturer benefitted from encapsulating a competitor's FDA-approved tablet in a gelatin capsule [22]. A better, simpler, and more common option is to implement a "double-dummy." Each active agent has a placebo identical to it. Each study participant would then take two medications. A pharmaceutical sponsor may sometimes have problems finding a matching placebo for a competitor's product.

If two or more active agents are being compared against placebo, it may not be feasible to make all drugs appear identical. As long as each active agent is not being compared against another, but only against placebo, one option is to create a placebo for each active drug or a so-called "double-dummy." Another option is to limit the number of placebos. For example, assume the trial consists of active drugs A, B, and C and placebo groups. If each group is of the same size, one-third of placebo groups will take a placebo designed to look like active drug A, one-third will take a placebo designed to look like drug B, and one-third, like active drug C. This design was successfully implemented in at least one reported study [23].

Coding of Drugs

By drug coding is meant the labeling of individual drug bottles or vials so that the identity of the drug is not disclosed. Coding is usually done by means of assigning a random set of numbers to the active drug and a different set to the control. As many different drug codes as are logistically feasible should be used. At least in smaller studies, each participant should have a unique drug code which remains with him for the duration of the trial. If only one code were used for each study group, unblinding a single participant would result in unblinding everybody. Furthermore, many drugs have specific side effects. One side effect in one participant may not be attributable to the drug, but a constellation of several side effects in several participants with the same drug code may easily unblind the whole study.

Unfortunately, in large studies, it becomes cumbersome logistically to make up and stock drugs under hundreds or thousands of unique codes. In multicenter trials, all participants at each clinic ought to have a unique code, if possible. Bar coding of the bottles with study medication is getting more common. This type of coding has no operational limits on the number of unique codes; it simplifies keeping an accurate and current inventory of all study medications and helps assure that each participant is dispensed his assigned study medication.

Official Unblinding

A procedure should be developed to break the blind quickly for any individual participant at any time should it be in his best interest. Such systems include having

labels on file in the hospital pharmacy or other accessible locations, or having an "on call" 24 hour-a-day process so that the assignment can be decoded. In order to avoid needless breaking of the code, someone other than the investigator could hold a list that reveals the identity of each drug code. Alternatively, each study medication bottle label might have a sealed tear-off portion that would be filed in the pharmacy or with the participant's records. In an emergency, the seal could be opened and the drug identity revealed. Care should be taken to ensure that the sealed portion is of appropriate color and thickness to prevent reading through it. In one study, the sealed labels attached to the medication bottles were transparent when held up to strong light.

Official breaking of the blind may be necessary. There are bound to be situations that require disclosures, especially in long-term studies. Perhaps the study drug requires tapering the dosage. In an emergency, knowledge that a participant is or is not on the active drug would indicate whether tapering is necessary. Children may get hold of study pills and swallow them. Usually, most emergencies can be handled by withdrawing the medication without breaking the blind. When the treating physician is different from the study investigator, a third party can obtain the blinded information from the pharmacy or central data repository and relate the information to the treating physician. In this way, the participant and the study investigator need not be unblinded. Knowledge of the kind of intervention seldom influences emergency care of the participant, and such reviews have helped reduce the frequency of further unblinding. When unblinding does occur, the investigator should review and report the circumstances which led to it in the results paper.

In summary, double-blind trials require careful planning and constant monitoring to ensure that the blind is maintained and that participant safety is not jeopardized.

Inadvertent Unblinding

The phrase "truly double-blind study" was used earlier. While many studies are designed as double- or single-blind, it is unclear how many, in fact, are truly and completely blind. Intended physiologic effects of a drug may be readily observable. Moreover, drugs have side effects, some of which are fairly characteristic. Known pharmaceutical effects of the study medication may lead to unblinding. Inhalation of short-acting beta-agonists causes tremor and tachycardia within minutes in most users. Even the salt of the active agent can cause side effects that lead to unblinding. For example, the blinded design was broken in a clinical trial comparing the commonly used ranitidine hydrochloride to a new formulation of ranitidine bismuth citrate. The bismuth-containing compound colored the tongue of its users black. Rifampin, a standard treatment for tuberculosis, causes the urine to change color. Existence or absence of such side effects does not necessarily unblind drug assignment since all people on drugs do not develop reactions and some people on placebo develop events which can be mistaken for drug side effects. It is well known that aspirin is associated with gastrointestinal problems. In the Women's Health Study

[24], 2.7% of the participants in the low-aspirin group developed peptic ulcer. On the other hand, 2.1% of the placebo participants had the same condition. This difference is highly significant ($p < 0.001$), but having an ulcer, in itself, would not unblind.

Occasionally, accidental unblinding occurs. In some studies, a special center labels and distributes drugs to the clinic where participants are seen. Obviously, each carton of drugs sent from the pharmaceutical company to this distribution center must contain a packing slip identifying the drug. The distribution center puts coded labels on each bottle and removes the packing slip before sending the drugs to the investigator. In one instance, one carton contained two packing slips by mistake. The distribution center, not realizing this, shipped the carton to the investigator with the second packing slip enclosed. Thus, it is advisable to empty cartons completely before re-using them.

Laboratory errors have also occurred. These are particularly likely when, to prevent unblinding, only some laboratory results are given to the investigators. Occasionally, investigators have received the complete set of laboratory results. This usually happens at the beginning of a study before "bugs" have been worked out, or when the laboratory hires new personnel who are unfamiliar with the procedures. If a commercial laboratory performs the study determinations, the tests should be done in a special area of the laboratory, with safeguards to prevent study results from getting intermingled with routine work. Routine laboratory panels obtained during regular clinical care of patients may include laboratory results that could lead to unblinding. In a large, long-term trial of a lipid-lowering drug, the investigators were discouraged from getting serum cholesterol determination on their coronary patients. It is difficult to know how many complied.

In addition, monitoring the use of study medication prescribed outside the study is essential. Any group differences might be evidence of a deficiency in the blind. Another way of estimating the success of a double-blind design is to monitor specific intermediate effects of the study medication. The use of platelet aggregation in the Aspirin Myocardial Infarction Study is an example. An unusually large number of participants with non-aggregating platelets in the placebo group would raise the suspicion that the blind had been broken.

Assessment and Reporting of Blindness

The importance of blindness in avoiding bias is well established in clinical trials. However, the assessment and reporting of blindness do not always receive proper attention. Readers of trial reports are often given incomplete information about the type of blinding and its success during the trial. This is a potential concern since randomized trials with inadequate blinding, on average, show larger treatment effects than properly blinded trials [25].

In their systematic review of 819 articles of blinded randomized trials assessing pharmacologic treatment, Boutron et al. [26] considered three blinding methods – (1)

the initial blinding of participants and investigators, (2) the maintenance of this blinding, and (3) the blinding of those assessing trial outcomes. Overall, only 472 of the blinded reports (58%) described the method of blinding, while 13% gave some information, and 29% none at all. The methods to establish blinding were presented in 41% of the reports. These included different types of matching, the use of a "double-dummy" procedure, sham interventions, and masking of the specific taste of the active treatments. The methods to maintain blinding during the trial, reported in only 3% of the articles, included a blinded centralized system for dosage adjustments in all study groups and centralized assessment of intermediate treatment effects (i.e., lipid determinations in statin trial). The methods for blinded assessment were described in 14% of the reports. They are especially useful in trials when blinding of intervention cannot be established. The main method was a centralized assessment of the primary outcome by blinded classification committees.

In assessing the success of blindness, decisions have to be made when to conduct the assessment and what questions to ask. There are different views as to when to assess blindness – early after randomization, throughout the trial or at the end [27]. Early assessment in a double-blind trial would be a measure of the initial success of blinding. Repeated questioning may trigger the curiosity of the study participants. Assessment at the end "confounds failures in blinding with successes in pre-trial hunches about efficacy" [28]. If study participants do well, there is a tendency for them to predict that they received active treatment; if they have suffered events or perceived no improvement, their prediction is more likely to be placebo. Similarly, investigators' hunches about efficacy can also be influenced by their preconceived expectations as illustrated by Sackett [29]. He concluded that "We neither can nor need to test for blindness during and after trial,…" It is more important to look for the consequences that may result from its loss. These include controls getting the active study medication, compensatory co-intervention, and/or evidence that the participants on the "better" treatment downplayed symptoms and denied mild events. Even investigators may downplay "soft" outcomes to fit their hopes or expectations.

A review of trials assessing blinding and reporting the methods used, concluded that there are diverse views regarding what questions to ask [27]. Some investigators allowed the subjects to express uncertainty while others forced them to guess. In some studies, the participants were asked to express their certainty of the guess or to explain the reason(s) for their guesses. It has been pointed out that the different ways of asking the questions "will lead to different results that may not be directly comparable" [30].

In a survey of 191 placebo-controlled double-blind trials published in 1998–2001, the authors evaluated how often the success of blindness was reported [19]. Only 15 (8%) reported evidence of success, and of these 15 trials, blinding was imperfect in nine. A similar survey of 1,599 blinded randomized trials from 2001 reported that only 2% of the trials reported tests for the success of blinding [31]. Interestingly, many investigators had conducted, but not published such tests.

In 2010, the Consolidated Standards for Reporting of Trials (CONSORT) Group published a revised statement. Item 11 of the recommendations asks two questions

about blinding: “If done, who was blinded after assignment of interventions (for example, participants, care providers, those assessing outcomes) and how? If relevant, description of the similarity of interventions” [32]. We believe that monitoring and reporting of the success of blindness are important for two reasons. First, knowing that the level of this success is one of the measures of trial quality may keep the investigators’ attention on blindness during trial conduct. Second, it helps readers of the trial report determine how much confidence they can place on the trial results.

References

1. Mausner JS, Bahn AK. *Epidemiology: An Introductory Text.* Philadelphia: W.B. Saunders, 1974.
2. Devereaux PJ, Manns BJ, Ghali WA, et al. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. *JAMA* 2001;285:2000–2003.
3. Park J, White AR, Stevinson C, Ernst E. Who are we blinding? A systematic review of blinded clinical trials. *Perfusion* 2001;14:296–304.
4. Hansson L, Hedner T, Dahlöf B. Prospective randomized open blinded end-point (PROBE) study. A novel design for intervention trials. Prospective Randomized Open Blinded End-Point. *Blood Press* 1992;1:113–119.
5. The Women’s Health Initiative Study Group. Design of the Women’s Health Initiative clinical trial and observational study. *Control Clin Trials* 1998;19:61–109.
6. Karlowski TR, Chalmers TC, Frenkel LD, et al. Ascorbic acid for the common cold. A prophylactic and therapeutic trial. *JAMA* 1975;231:1038–1042.
7. Lewis TL, Karlowski TR, Kapikian AZ, et al. A controlled clinical trial of ascorbic acid for the common cold. *Ann N Y Acad Sci* 1975;258:505–512.
8. European Coronary Surgery Study Group. Coronary-artery bypass surgery in stable angina pectoris: survival at two years. *Lancet* 1979;i:889–893.
9. Schechter PJ, Friedewald WT, Bronzert DA, et al. Idiopathic hypogeusia: a description of the syndrome and a single-blind study with zinc sulfate. *Int Rev Neurobiol* 1972;(suppl 1):125–140.
10. Henkin RI, Schechter PJ, Friedewald WT, et al. A double blind study of the effects of zinc sulfate on taste and smell dysfunction. *Am J Med Sci* 1976;272:285–299.
11. Halperin JL, and The Executive Steering Committee on behalf of the SPORTIF III and V Study Investigators. Ximelagatran compared with warfarin for prevention of thromboembolism in patients with nonvalvular atrial fibrillation: rationale, objectives, and design of a pair of clinical studies and baseline patient characteristics (SPORTIF III and V). *Am Heart J* 2003;146:431–438.
12. Tashkin DP, Celli B, Senn S, et al. for the UPLIFT Study Investigators. A 4-year trial of tiotropium in chronic obstructive pulmonary disease. *N Engl J Med* 2008;359:1543–1554.
13. Hertzberg V, Chimowitz M, Lynn M, et al. Use of dose modification schedules is effective for blinding trials of warfarin: evidence from the WASID study. *Clin Trials* 2008;5:25–30.
14. Coumadin Aspirin Reinfarction Study (CARS) Investigators. Randomised double-blind trial of fixed low-dose warfarin with aspirin after myocardial infarction. *Lancet* 2008; 350:389–396.
15. Meinert CL. Masked monitoring in clinical trials – blind stupidity? (Sounding Board). *N Engl J Med* 1998;338:1381–1382.
16. Wittes J, Boissel J-P, Furberg CD, et al. Stopping the Randomized Aldactone Evaluation Study early for efficacy. In DeMets DL, Furberg CD, Friedman LM, (eds.). *Data Managing in Clinical Trials. A Case Study Approach.* New York: Springer, 2006, pp. 148–157.
17. Aspirin Myocardial Infarction Study Research Group. A randomized, controlled trial of aspirin in persons recovered from myocardial infarction. *JAMA* 1980;243:661–669.

18. Hill LE, Nunn AJ, Fox W. Matching quality of agents employed in “double-blind” controlled clinical trials. *Lancet* 1976;i:352–356.
19. Fergusson D, Glass KC, Waring D, Shapiro S. Turning a blind eye: the success of blinding reported in a random sample of randomised, placebo controlled trials. *Br Med J* 2004;328:432.
20. Farr BM, Gwaltney JM Jr. The problems of taste in placebo matching: an evaluation of zinc gluconate for the common cold. *J Chronic Dis* 1987;40:875–879.
21. DeKosky ST, Williamson JD, Fitzpatrick AL, et al. for the Ginkgo Evaluation of Memory (GEM) Study Investigators. *Ginkgo biloba* for prevention of dementia. A randomized controlled trial. *JAMA* 2008;300:2253–2262.
22. Mathew NT, Schoenen J, Winner P, et al. Comparative efficacy of eletriptan 40 mg versus sumatriptan 100 mg. *Headache* 2003;43:214–222.
23. The CAPS Investigators. The Cardiac Arrhythmia Pilot Study. *Am J Cardiol* 1986;57:91–95.
24. Ridker PM, Cook NP, Lee I-M, et al. A randomized trial of low-dose aspirin in the primary prevention of cardiovascular disease in women. *N Engl J Med* 2005;352:1293–1304.
25. Schulz KF, Chalmers I, Hayes R, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408–412.
26. Boutron I, Estellat C, Guittet L, et al. Methods of blinding in reports of randomized controlled trials assessing pharmacologic treatments: a systematic review. *PLoS Med* 2006;3:1931–1939.
27. Boutron I, Estellat C, Ravaud P. A review of blinding in randomized controlled trials found results inconsistent and questionable. *J Clin Epidemiol* 2005;58:1220–1226.
28. Sackett DL. Turning a blind eye. Why we don’t test for blindness at the end of our trials. (Letter) *Br Med J*, doi:10:1136/bmj.328.7448.1136-a (published 8 May 2004).
29. Sackett DL. Commentary: Measuring the success of blinding in RCTs: don’t, must, can’t or needn’t? *Int J Epidemiol* 2007;36:664–665.
30. Desbiens NA. Lessons learned from attempts to establish the blind in placebo-controlled trials of zinc for the common cold (Editorial). *Ann Intern Med* 2000;133:302–303.
31. Hróbjartsson A, Forfang E, Haahr MT, et al. Blinded trials taken to the test: an analysis of randomized clinical trials that report tests for the success of blinding. *Int J Epidemiol* 2007;36:654–663.
32. Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: Guidelines for reporting parallel group randomised trials. *PLoS Med* 2010; 7: e1000251.

Chapter 8
Sample Size

The size of the study should be considered early in the planning phase. In some instances, no formal sample size is ever calculated. Instead, the number of participants available to the investigators during some period of time determines the size of the study. Many clinical trials that do not carefully consider the sample size requirements turn out to lack the statistical power or ability to detect intervention effects of a magnitude that has clinical importance. In 1978, Freiman and colleagues [1] reviewed the power of 71 published randomized controlled clinical trials, which failed to find significant differences between groups. "Sixty-seven of the trials had a greater than 10% risk of missing a true 25% therapeutic improvement, and with the same risk, 50 of the trials could have missed a 50% improvement." In other instances, the sample size estimation may assume an unrealistically large intervention effect. Thus, the power for more realistic intervention effects will be low or less than desired. The danger in studies with low statistical power is that interventions that could be beneficial are discarded without adequate testing and may never be considered again. Certainly, many studies do contain appropriate sample size estimates, but many are still too small.

This chapter presents an overview of sample size estimation with some details. Several general discussions of sample size can be found elsewhere [2–11]. For example, Lachin [2] and Donner [7] have each written a more technical discussion of this topic. For most of the chapters, the focus is on sample size where the study is randomizing individuals. In some sections, the concept of sample size for randomizing clusters of individuals or organs within individuals is presented.

Fundamental Point

Clinical trials should have sufficient statistical power to detect differences between groups considered to be of clinical importance. Therefore, calculation of sample size with provision for adequate levels of significance and power is an essential part of planning.

Before a discussion of sample size and power calculations, it must be emphasized that, for several reasons, a sample size calculation provides only an estimate of the

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_8, © Springer Science+Business Media, LLC 2010

needed size of a trial [3]. First, parameters used in the calculation are estimates, and as such, have an element of uncertainty. Often these estimates are based on small studies. Second, the estimate of the relative effectiveness of the intervention over the control may be based on a population different from that intended to be studied. Third, the effectiveness is often overestimated since published pilot studies may be highly selected and researchers are often too optimistic. Fourth, during the final planning stage of a trial, revisions of inclusion and exclusion criteria may influence the types of participants entering the trial and thus alter earlier assumptions used in the sample size calculation. Assessing the impact of such changes in criteria and the screening effect is usually quite difficult. Trial experience indicates that participants enrolled into control groups usually do better than the population from which the participants were drawn. The reasons are not entirely clear. One factor could be that participants with the highest risk of developing the outcome of interest are excluded in the screening process. In trials involving chronic diseases, because of the research protocol, participants might receive more care and attention than they would normally be given, thus improving their prognosis. Participants assigned to the control group may, therefore, be better off than if they had not been in the trial at all. Finally, sample size calculations are based on mathematical models that may only approximate the true, but unknown, distribution of the response variables.

Due to the approximate nature of sample size calculations, the investigator should be as conservative as can be justified while still being realistic in estimating the parameters used in the calculation. If a sample size is drastically overestimated, the trial may be judged as unfeasible. If the sample size is underestimated, there is a good chance the trial will fall short of demonstrating any differences between study groups. In general, as long as the calculated sample size is realistically obtainable, it is better to overestimate the size and possibly terminate the trial earlier (Chap. 16) than to underestimate, and need to justify an increase in sample size or an extension in follow-up, or worse, to arrive at incorrect conclusions.

Statistical Concepts

An understanding of the basic statistical concepts of hypothesis testing, significance level, and power is essential for a discussion of sample size estimation. A brief review of these concepts follows. Further discussion can be found in many basic medical statistics textbooks [12–19] as well as selected review papers [4, 5, 7]. Those with no prior exposure to these basic statistical concepts might find these resources helpful.

Except where indicated, trials of one intervention group and one control group will be discussed. With some adjustments, sample size calculations can be made for studies with more than two groups. For example, in the Coronary Drug Study (CDP), five active intervention arms were each compared against one control arm [20]. Using the method of Dunnett [21], where the control group has the number of participants equal to the square root of the number assigned to the combined number

in the active intervention groups, the optimal size of the control arm in the CDP was determined to be 2.24 times the size of each individual active intervention arm [20]. In fact, the CDP used a factor of 2.5 in order to minimize variance. Other approaches are to use the Bonferroni adjustment to the alpha level [22]; that is, divide the overall alpha level by the number of comparisons, and use that revised alpha level in the sample size comparison.

Before computing sample size, the primary response variable used to judge the effectiveness of intervention must be identified (see Chap. 3). This chapter will consider sample size estimation for three basic kinds of outcomes: [1] dichotomous response variables, such as success and failure, [2] continuous response variables, such as blood pressure level or a change in blood pressure, and [3] time to failure (or occurrence of a clinical event).

For the dichotomous response variables, the event rates in the intervention group (p_I) and the control group (p_C) are compared. For continuous response variables, the true, but unknown, mean level in the intervention group (μ_I) is compared with the mean level in the control group (μ_C). For survival data, a hazard rate, λ, is often compared for the two study groups or at least is used for sample size estimation. Sample size estimates for response variables which do not exactly fall into any of the three categories can usually be approximated by one of them.

In terms of the primary response variable, p_I will be compared with p_C or μ_I will be compared with μ_C. This discussion will use only the event rates, p_I, and p_C, although the same concepts will hold if response levels μ_I and μ_C are substituted appropriately. Of course, the investigator does not know the true values of the event rates. The clinical trial will give him only estimates of the event rates, p_I and p_C. Typically, an investigator tests whether or not a true difference exists between the event rates of participants in the two groups. The traditional way of indicating this is in terms of a null hypothesis, denoted H_0, which states that no difference between the true event rates exists (H_0: $p_C - p_I = 0$). The goal is to test H_0 and decide whether or not to reject it. That is, the null hypothesis is assumed to be true until proven otherwise.

Because only estimates of the true event rates are obtained, it is possible that, even if the null hypothesis is true ($p_C - p_I = 0$), the observed event rates might by chance be different. If the observed differences in event rates are large enough by chance alone, the investigator might reject the null hypothesis incorrectly. This false positive finding, or *Type I error*, should be made as few times as possible. The probability of this Type I error is called the significance level and is denoted by α. The probability of observing differences as large as, or larger than the difference actually observed given that H_0 is true is called the "p value," denoted as p. The decision will be to reject H_0 if $p \le \alpha$. While the chosen level of α is somewhat arbitrary, the ones used and accepted traditionally are 0.01, 0.025, or 0.05. As will be shown later, as α is set smaller, the required sample size estimate increases.

If the null hypothesis is not in fact true, then another hypothesis, called the alternative hypothesis, denoted by H_A, must be true. That is, the true difference between the event rates p_I and p_C is some value δ where $\delta \neq 0$. The observed difference, $\hat{p}_C - \hat{p}_I$, can be quite small by chance alone even if the alternative hypothesis is true.

Therefore, the investigator could, on the basis of small observed differences, fail to reject H_0 when he should. This is called a *Type II error*, or a false negative result. The probability of a Type II error is denoted by β. The value of β is dependent on the specific value of δ, the true but unknown difference in event rates between the two groups, as well as on the sample size and α. The probability of correctly rejecting H_0 is denoted by $1-\beta$ and is called the power of the study. Power quantifies the ability of the study to find true differences of various values δ. Since β is a function of α, the sample size and δ, $1-\beta$ is also a function of these parameters. The plot of $1-\beta$ versus δ for a given sample size is called the power curve and is depicted in Fig. 8.1. On the horizontal axis, values of δ are plotted from 0 to an upper value, δ_A (0.25 in this figure). On the vertical axis, the probability or power of detecting a true difference δ is shown for a given significance level and sample size. In constructing this specific power curve, a sample size of 100 in each group, a one-sided significance level of 0.05 and a control group event rate of 0.5 (50%) were assumed. Note that as δ increases, the power to detect δ also increases. For example, if δ=0.10, the power is approximately 0.40. When δ=0.20, the power increases to about 0.90. Typically, investigators like to have a power $(1-\beta)$ of at least 0.80, but often around 0.90 or 0.95 when planning a study; that is to have an 80, 90, or 95% chance of finding a statistically significant difference between the event rates, given that a difference, δ, actually exists.

Since the significance level α should be small, say 0.05 or 0.01, and the power $(1-\beta)$ should be large, say 0.90 or 0.95, the only quantities which are left to vary are δ, the size of the difference being tested for, and the total sample size. In planning a clinical trial, the investigator hopes to detect a difference of specified

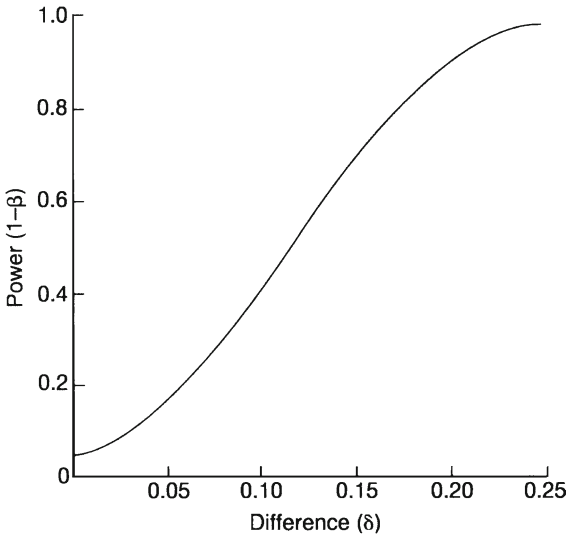


Fig. 8.1 A power curve for increasing differences (δ) between the control group rate of 0.5 and the intervention group rate with a one-sided significance level of 0.05 and a total sample size (2*N*) of 200

magnitude δ or larger. One factor that enters into the selection of δ is the minimum difference between groups that is judged to be clinically important. In addition, previous research may provide estimates of δ. This is part of the question being tested as discussed in Chap. 3. The exact nature of the calculation of the sample size, given α, $1-\beta$, and δ is considered here. It can be assumed that the randomization strategy will allocate an equal number (N) of participants to each group, since the variability in the responses for the two groups is approximately the same, equal allocation provides a slightly more powerful design than does unequal allocation. For unequal allocation to yield an appreciable increase in power, the variability needs to be substantially different in the groups [23]. Since equal allocation is usually easier to implement, it is the more frequently used strategy and will be assumed here for simplicity.

Before a sample size can be calculated, classical statistical theory says that the investigator must decide whether he is interested in differences in one direction only (one-sided test) – say improvements in intervention over control – or in differences in either direction (two-sided test). This latter case would represent testing the hypothesis that the new intervention is either better or worse than the control. In general, two-sided tests should be used unless there is a very strong justification for expecting a difference in only one direction. An investigator should always keep in mind that any new intervention could be harmful as well as helpful. However, as discussed in Chap. 16, some investigators may not be willing to prove the intervention harmful and would terminate a study if the results are suggestive of harm. A classic example of this issue was provided by the Cardiac Arrhythmia Suppression Trial or CAST [24]. This trial was initially designed as a one-sided, 0.025 significance level hypothesis test that anti-arrhythmic drug therapy would reduce the incidence of sudden cardiac death. Since the drugs were already marketed, harmful effects were not expected. Despite the one-sided hypothesis in the design, the monitoring process used a two-sided, 0.05 significance level approach. In this respect, the level of evidence for benefit was the same for either the one-sided 0.025 or two-sided 0.05 significance level design. As it turned out, the trial was terminated early due to increased mortality in the intervention group (see Chap. 16).

If a one-sided test of hypothesis is chosen, in most circumstances, the significance level ought to be half what the investigator would use for a two-sided test. For example, if 0.05 is the two-sided significance level typically used, 0.025 would be used for the one-sided test. As done in the CAST trial, this requires the same degree of evidence or scientific documentation to declare a treatment effective, regardless of the one-sided versus two-sided question. In this circumstance, a test for negative or harmful effects might also be done at the 0.025 level. This in effect, provides two one-sided 0.025 hypothesis tests for an overall 0.05 significance level.

As mentioned above, the total sample size $2N$ (N per arm) is a function of the significance level (α), the power ($1-\beta$), and the size of the difference in response (δ), which is to be detected. Changing either α, $1-\beta$, or δ will result in a change in $2N$. As the magnitude of the difference δ decreases, the larger the sample size must be to guarantee a high probability of finding that difference. If the calculated

sample size is larger than can be realistically obtained, then one or more of the parameters in the design may need to be reconsidered. Since the significance level is usually fixed at 0.05, 0.025, or 0.01, the investigator should generally reconsider the value selected for δ and increase it, or keep δ the same and settle for a less powerful study. If neither of these alternatives is satisfactory, serious consideration should be given to abandoning the trial.

Rothman [25] argued that journals should encourage using confidence intervals to report clinical trial results instead of significance levels. Several researchers [14, 25, 26] discuss sample size formulas from this approach. Confidence intervals are constructed by computing the observed difference in event rates and then adding and subtracting a constant times the standard error of the difference. This provides an interval surrounding the observed estimated difference obtained from the trial. The constant is determined so as to give the confidence interval the correct probability of including the true, but unknown difference. This constant is related directly to the critical value used to evaluate test statistics. Trials often use a two-sided α level test (e.g., α=0.05) and a corresponding $(1-\alpha)$ confidence interval (e.g., 95%). If the $1-\alpha$ confidence interval excludes zero or no difference, we would conclude that the intervention has an effect. If the interval contains zero difference, no intervention effect would be claimed. However, differences of importance could exist, but might not be detected or not be statistically significant because the sample size was too small. For testing the null hypothesis of no treatment effect, hypothesis testing and confidence intervals give the same conclusions. However, confidence intervals provide more information on the range of the likely difference that might exist. For sample size calculations, the desired confidence interval width must be specified. This may be determined, for example, by the smallest difference between two event rates that would be clinically meaningful and important. Under the null hypothesis of no treatment effect, half the desired interval width is equal to the difference specified in the alternative hypothesis. The sample size calculation methods presented here do not preclude the presentation of results as confidence intervals and, in fact, investigators ought to do so. However, unless there is an awareness of the relationship between the two approaches, as McHugh and Le [26] have pointed out, the confidence interval method might yield a power of only 50% to detect a specified difference. This can be seen later, when sample size calculations for comparing proportions are presented. Thus, some care needs to be taken in using this method.

So far, it has been assumed that the data will be analyzed only once at the end of the trial. However, as discussed in Chap. 16, the response variable data may be reviewed periodically during the course of a study. Thus, the probability of finding significant differences by chance alone is increased [27]. This means that the significance level α may need to be adjusted to compensate for the increase in the probability of a Type I error. For purposes of this discussion, we assume that α carries the usual values of 0.05, 0.025, or 0.01. The sample size calculation should also employ the statistic which will be used in data analysis. Thus, there are many sample size formulations. Methods that have proven useful will be discussed in the rest of this chapter.

Dichotomous Response Variables

We shall consider two cases for response variables which are dichotomous, that is, yes or no, success or failure, presence or absence. The first case assumes two independent groups or samples [28–40]. The second case is for dichotomous responses within an individual, or paired responses [41–45].

Two Independent Samples

Suppose the primary response variable is the occurrence of an event over some fixed period of time. The sample size calculation should be based on the specific test statistic that will be employed to compare the outcomes. The null hypothesis H_0 ($p_C - p_I = 0$) is compared to an alternative hypothesis H_A ($p_C - p_I \neq 0$). The estimates of p_I and p_C are $\hat{p}_I$ and $\hat{p}_C$ where $\hat{p}_I = r_I / N_I$ and $\hat{p}_C = r_C / N_C$ with r_I and r_C being the number of events in the intervention and control groups and N_I and N_C being the number of participants in each group. The usual test statistic for comparing such dichotomous or binomial responses is

$$Z = (\hat{p}_C - \hat{p}_I) / \sqrt{\overline{p}(1-\overline{p})(1/N_C + 1/N_I)}$$

where $\overline{p} = (r_I + r_C)/(N_I + N_C)$. The square of the Z statistic is algebraically equivalent to the chi-square statistic, which is often employed as well. For large values of N_I and N_C, the statistic Z has approximately a normal distribution with mean 0 and variance 1. If the test statistic Z is larger in absolute value than a constant Z_α, the investigator will reject H_0 in the two-sided test.

The constant Z_α is often referred to as the critical value. The probability of a standard normal random variable being larger in absolute value than Z_α is α. For a one-sided hypothesis, the constant Z_α is chosen such that the probability that Z is greater (or less) than Z_α is α. For a given α, Z_α is larger for a two-sided test than for a one-sided test (Table 8.1). Z_α for a two-sided test with $\alpha = 0.10$ has the same value as Z_α for a one-sided test with $\alpha = 0.05$. While a smaller sample size can be achieved with a one-sided test compared to a two-sided test at the same α level, we in general do not recommend this approach as discussed earlier.

The sample size required for the design to have a significance level α and a power of $1-\beta$ to detect true differences of at least δ between the event rates p_I and p_C can be expressed by the formula [2]:

$$2N = 2\left\{Z_\alpha\sqrt{2\overline{p}(1-\overline{p})} + Z_\beta\sqrt{p_C(1-p_C)+p_I(1-p_I)}\right\}^2 \Big/ (p_C - p_I)^2$$

where $2N$=total sample size (N participants/group) with $\overline{p} = (p_C + p_I)/2$; Z_α is the critical value which corresponds to the significance level α; and Z_β is the value of the standard normal value not exceeded with probability β. Z_β corresponds to the

Table 8.1 Z_α for sample size formulas for various values of α

	Z_α	
α	One-sided test	Two-sided test
0.10	1.282	1.645
0.05	1.645	1.960
0.025	1.960	2.240
0.01	2.326	2.576

power $1-\beta$ (e.g., if $1-\beta=0.90$, $Z_\beta=1.282$). Values of Z_α and Z_β are given in Tables 8.1 and 8.2 for several values of α and $1-\beta$. More complete tables may be found in most introductory texts [12–19]. Note that the definition of $\bar{p}$ given earlier is equivalent to the definition of $\bar{p}$ given here when $N_I=N_C$; that is, when the two study groups are of equal size. An alternative to the above formula is given by

$$2N = 4(Z_\alpha + Z_\beta)^2 \bar{p}(1-\bar{p})/(p_C - p_I)^2$$

These two formulas give approximately the same answer and either may be used for the typical clinical trial.

Example: Suppose the annual event rate in the control group is anticipated to be 20%. The investigator hopes that the intervention will reduce the rate to 15%. The study is planned so that each participant will be followed for 2 years. Therefore, if the assumptions are accurate, approximately 40% of the participants in the control group and 30% of the participants in the intervention group will develop an event. Thus, the investigator sets $p_C=0.40$, $p_I=0.30$, and therefore $\bar{p} = (0.4+0.3)/2 = 0.35$. The study is designed as two-sided with a 5% significance level and 90% power. From Tables 8.1 and 8.2, the two-sided 0.05 critical value is 1.96 for Z_β and 1.282 for Z_β. Substituting these values into the right-hand side of the first sample size formula yields $2N$ to be

$$2\left\{1.96\sqrt{2(0.35)(0.65)} + 1.282\sqrt{0.4(0.6)+0.3(0.7)}\right\}^2 \Big/ (0.4-0.3)^2$$

Evaluating this expression, $2N$ equals 952.3. Using the second formula, $2N$ is $4(1.96+1.202)^2\,(0.35)(0.65)/(0.4-0.3)^2$ or $2N=956$. Therefore, after rounding up to the nearest 10, the calculated total sample size by either formula is 960, or 480 in each group.

Sample size estimates using the first formula are given in Table 8.3 for a variety of values of p_I and p_C, for two-sided tests, and for $\alpha=0.01$, 0.025, and 0.05 and $1-\beta=0.80$ or 0.90. For the example just considered with $\alpha=0.05$ (two-sided), $1-\beta=0.90$, $p_C=0.4$, and $p_I=0.3$, the total sample size using Table 8.3 is 960. This table shows that, as the difference in rates between groups increases, the sample size decreases.

The event rate in the intervention group can be written as $p_I=(1-k)p_C$ where k represents the proportion that the control group event rate is expected to be reduced by the intervention. Figure 8.2 shows the total sample size $2N$ versus k for

Table 8.2 Z_β for sample size formulas for various values of power $(1-\beta)$

$1-\beta$	Z_β
0.50	0.00
0.60	0.25
0.70	0.53
0.80	0.84
0.85	1.036
0.90	1.282
0.95	1.645
0.975	1.960
0.99	2.326

Table 8.3 Sample size

		2α (Two-sided)					
Alpha/power		0.01		0.025		0.05	
p_C	p_1	0.90	0.80	0.90	0.80	0.90	0.80
0.6	0.5	1,470	1,160	1,230	940	1,040	780
	0.4	370	290	310	240	260	200
	0.3	160	130	140	110	120	90
	0.20	90	70	80	60	60	50
0.5	0.40	1,470	1,160	1,230	940	1,040	780
	0.30	360	280	300	230	250	190
	0.25	220	180	190	140	160	120
	0.20	150	120	130	100	110	80
0.4	0.30	1,360	1,060	1,130	870	960	720
	0.25	580	460	490	370	410	310
	0.20	310	250	260	200	220	170
0.3	0.20	1,120	880	930	710	790	590
	0.15	460	360	390	300	330	250
	0.10	240	190	200	150	170	130
0.20	0.15	3,440	2,700	2,870	2,200	2,430	1,810
	0.10	760	600	630	490	540	400
	0.05	290	230	240	190	200	150
0.10	0.05	1,650	1,300	1,380	1,060	1,170	870

several values of p_C using a two-sided test with $\alpha=0.05$ and $1-\beta=0.90$. In the example where $p_C=0.4$ and $p_1=0.3$, the intervention is expected to reduce the control rate by 25% or $k=0.25$. In Fig. 8.2, locate $k=0.25$ on the horizontal axis and move up vertically until the curve labeled $p_C=0.4$ is located. The point on this curve corresponds to a $2N$ of approximately 960. Notice that as the control group event rate p_C decreases, the sample size required to detect the same proportional reduction increases. Trials with small event rates (e.g., $p_C=0.1$) require large sample sizes unless the interventions have a dramatic effect.

In order to make use of the sample size formula or table, it is necessary to know something about p_C and k. The estimate for p_C is usually obtained from previous

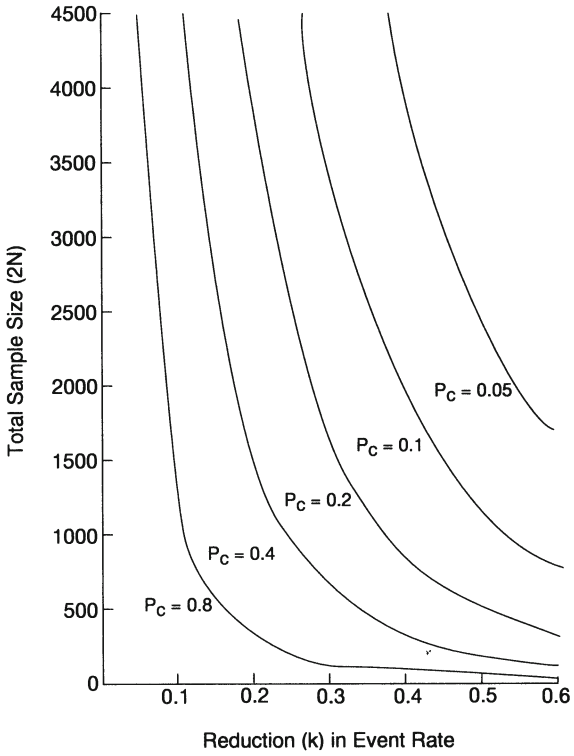


Fig. 8.2 Relationship between total sample size ($2N$) and reduction (k) in event rate for several control group event rates (p_C), with a two-sided significance level of 0.05 and power of 0.90

studies of similar people. In addition, the investigator must choose k based on preliminary evidence of the potential effectiveness of the intervention or be willing to specify some minimum difference or reduction that he wants to detect. Obtaining this information is difficult in many cases. Frequently, estimates may be based on a small amount of data. In such cases, several sample size calculations based on a range of estimates help to assess how sensitive the sample size is to the uncertain estimates of p_C, k, or both. The investigator may want to be conservative and take the largest, or nearly largest, estimate of sample size to be sure his study has sufficient power. The power $(1-\beta)$ for various values of δ can be compared for a given sample size $2N$, significance level α, and control rate p_C. By examining a power curve such as in Fig. 8.1, it can be seen what power the trial has for detecting various differences in rates, δ. If the power is high, say 0.80 or larger, for the range of values δ that are of interest, the sample size is probably adequate. The power curve can be especially helpful if the number of available participants is relatively fixed and the investigator wants to assess the probability that the trial can detect any of a variety of reductions in event rates.

Investigators often overestimate the number of eligible participants who can be enrolled in a trial. The actual number enrolled may fall short of goal. To examine

the effects of smaller sample sizes on the power of the trial, the investigator may find it useful to graph power as a function of various sample sizes. If the power falls far below 0.8 for a sample size that is very likely to be obtained, he can expand the recruitment effort, hope for a larger intervention effect than was originally assumed, accept the reduced power and its consequences or abandon the trial.

To determine the power, the second sample size equation in this section is solved for Z_β.

$$Z_\beta = \frac{-Z_\alpha\sqrt{2\overline{p}(1-\overline{p})} + \sqrt{N}(p_C - p_I)}{\sqrt{p_C(1-p_C) + p_1(1-p_I)}}$$

where $\overline{p}$ as before is $(p_C + p_1)/2$. The term Z_β can be translated into a power of $1-\beta$ by use of Table 8.2. For example, let $p_C = 0.4$ and $p_1 = 0.3$. For a significance level of 0.05 in a two-sided test of hypothesis, Z_α is 1.96. In a previous example, it was shown that a total sample of approximately 960 participants or 480 per group is necessary to achieve a power of 0.90. Substituting $Z_\alpha = 1.96$, $N = 480$, $p_C = 0.4$, and $p_1 = 0.3$, a value for $Z_\beta = 1.295$ is obtained. The closest value of Z_β in Table 8.2 is 1.282 which corresponds to a power of 0.90. (If the exact value of $N = 476$ were used, the value of Z_β would be 1.282.) Suppose an investigator thought he could get only 350 participants per group instead of the estimated 480. Then, $Z_\beta = 0.818$, which means that the power is somewhat less than 0.80. If the value of Z_β is negative, the power is less than 0.50. For more details of power calculations, a standard text in biostatistics [12–19] should be consulted.

For a given $2N$, α, $1-\beta$, and p_C, the reduction in event rate that can be detected can also be calculated. This function is nonlinear, and therefore the details will not be presented here. Approximate results can be obtained by scanning Table 8.3, by using the calculations for several p_1 until the sample size approaches the planned number, or by using a figure where sample sizes have been plotted. In Fig. 8.2, α is 0.05 and $1-\beta$ is 0.90. If the sample size is selected as 1,000, with $p_C = 0.4$, k is determined to be about 0.25. This means that the expected p_1 would be 0.3. As can be seen in Table 8.3, the actual sample size for these assumptions is 960.

The above approach yields an estimate which is more accurate as the sample size increases. Modifications [28–35, 37] have been developed which give some improvement in accuracy to the approximate formula presented for small studies. However, given that sample size estimation is somewhat imprecise due to assumptions of intervention effects and event rates, the formulation presented is probably adequate for most clinical trials.

Designing a trial comparing proportions using the confidence interval approach, we would need to make a series of assumptions as well [3, 23, 37]. A $100(1-\alpha)\%$ confidence interval for a treatment comparison θ would be of the general form $\hat{\theta} \pm Z_\alpha SE(\hat{\theta})$, where $\hat{\theta}$ is the estimate for θ and $SE(\hat{\theta})$ is the standard error of $\hat{\theta}$. In this case, the specific form would be:

$$(\hat{p}_I - \hat{p}_C) \pm Z_\alpha\sqrt{\overline{p}(1-\overline{p})(1/N_I + 1/N_C)}$$

If we want the width of the confidence interval (CI) not to exceed W_{CI}, where W_{CI} is the difference between the upper confidence limit and the lower confidence limit, then if $N=N_I=N_C$, the width W_{CI} can be expressed simply as:

$$W_{CI} = 2Z_\alpha\sqrt{\overline{p}(1-\overline{p})(2/N)}$$

or after solving this equation for N,

$$N = \frac{8Z_a^2\overline{p}(1-\overline{p})}{(W_{CI})^2}$$

Thus, if α is 0.05 for a 95% confidence interval, $p_C=0.4$ and $p_I=0.3$ or 0.35, $N=8(1.96)^2(0.35)(0.65)/(W_{CI})^2$. If we desire the upper limit of the confidence interval to be not more than 0.10 from the estimate or the width to be twice that, then $W_{CI}=0.20$ and $N=175$ or $2N=350$. Notice that even though we are essentially looking for differences in p_C-p_I to be the same as our previous calculation, the sample size is smaller. If we let $p_C-p_I=W_{CI}/2$ and substitute this into the previous sample size formula, we obtain

$$\begin{aligned} 2N &= 2\left\{Z_\alpha + Z_\beta\right\}^2 \overline{p}(1-\overline{p})/(W_{CI}/2)^2 \\ &= 8\left\{Z_\alpha + Z_\beta\right\}^2 \overline{p}(1-\overline{p})/(W_{CI})^2 \end{aligned}$$

This formula is very close to the confidence interval formula for two proportions. If we select 50% power, β is 0.50 and Z_β is 0 which would yield the confidence interval formula. Thus, a confidence interval approach gives 50% power to detect differences of $W_{CI}/2$. This may not be adequate, depending on the situation. In general, we prefer to specify greater power (e.g., 80–90%) and use the previous approach.

Analogous sample size estimation using the confidence interval approach may be used for comparing means, hazard rates, or regression slopes. We do not present details of these since we prefer to use designs which yield power greater than that obtained from a confidence interval approach.

Paired Dichotomous Response

For designing a trial where the paired outcomes are binary, the sample size estimate is based on McNemar's test [41–45]. We want to compare the frequency of success within an individual on intervention with the frequency of success on control (i.e., p_I-p_C). McNemar's test compares difference in discordant responses within an individual p_I-p_C, between intervention and control.

In this case, the number of paired observations, N_p, may be estimated by:

$$N_p = \left[Z_\alpha\sqrt{f} + Z_\beta\sqrt{f-d^2}\right]^2 \Big/ d^2$$

where d = difference in the proportion of successes ($d = p_I - p_C$) and f is the proportion of participants whose response is discordant. An alternative approximate formula for N_p is

$$N_p = \frac{[Z_\alpha + Z_\beta]^2 f}{d^2}$$

Example: Consider an eye study where one eye is treated for loss in visual acuity by a new laser procedure and the other eye is treated by standard therapy. The failure rate on the control P_C is estimated to be 0.40, and the new procedure is projected to reduce the failure rate to 0.20. The discordant rate f is assumed to be 0.50. Using the latter sample size formula for a two-sided 5% significance level and 90% power, the number of pairs N_p is estimated as 132. If the discordant rate is 0.8, then 210 pairs of eyes will be needed.

Adjusting Sample Size to Compensate for Nonadherence

During the course of a clinical trial, participants will not always adhere to their prescribed intervention schedule. The reason is often that the participant cannot tolerate the dosage of the drug or the degree of intervention prescribed in the protocol. The investigator or the participant may then decide to follow the protocol with less intensity. At all times during the conduct of a trial, the participant's welfare must come first and meeting those needs may not allow some aspects of the protocol to be followed. Planners of clinical trials must recognize this possibility and attempt to account for it in their design. Examples of adjusting for nonadherence with dichotomous outcomes can be found in several clinical trials [46–53].

In the intervention group, a participant who does not adhere to the intervention schedule is often referred to as a "drop-out." Participants who stop the intervention regimen lose whatever potential benefit the intervention might offer. Similarly, a participant on the control regimen may at some time begin to use the intervention that is being evaluated. This participant is referred to as a "drop-in." In the case of a drop-in, a physician may decide, for example, that surgery is required for a participant assigned to medical treatment in a clinical trial of surgery versus medical care [50]. Drop-in participants from the control group who start the intervention regimen will receive whatever potential benefit or harm that the intervention might offer. Therefore, both the drop-out and drop-in participants must be acknowledged because they tend to dilute any difference between the two groups which might be produced by the intervention. This simple model does not take into account the situation in which one level of an intervention is compared to another level of the intervention. More complicated models for nonadherence adjustment can be developed. Regardless of the model, it must be emphasized that the assumed event rates in the control and intervention groups are modified by participants who do not adhere to the study protocol.

People who do not adhere should remain in the assigned study groups and be included in the analysis. The rationale for this is discussed in Chap. 17. The basic point to be made here is that eliminating participants from analysis or transferring participants to the other group could easily bias the results of the study. However, the observed δ is likely to be less than projected because of nonadherence and thus have an impact on the power of the clinical trial. A reduced δ, of course, means that either the sample size must be increased or the study will have smaller power than intended. Lachin [2] has proposed a simple formula to adjust crudely the sample size for a drop-out rate of proportion R_O. This can be generalized to adjust for drop-in rates, R_I, as well. The unadjusted sample size N should be multiplied by the factor $\{1/(1-R_O-R_I)\}^2$ to get the adjusted sample size per arm, N^*. Thus, if $R_O=0.20$ and $R_I=0.05$, the originally calculated sample should be multiplied by $1/(0.75)^2$ or 16/9, and increased by 78%. This formula gives some quantitative idea of the effect of drop-out on the sample size.

$$N^* = N/(1-R_O-R_I)^2$$

However, more refined models to adjust sample sizes for drop-outs from the intervention to the control [54–60] and for drop-ins from the control to the intervention regimen [56] have been developed. They adjust for the resulting changes in p_I and p_C, the adjusted rates being denoted p_I^* and p_C^*. These models also allow for another important factor, which is the time required for the intervention to achieve maximum effectiveness. For example, an anti-platelet drug may have an immediate effect; conversely, even though a cholesterol-lowering drug reduces serum levels quickly, it may require years to produce a maximum effect on coronary mortality.

Example: A drug trial [48] in post myocardial infarction participants illustrates the effect of drop-outs and drop-ins on sample size. In this trial, total mortality over a 3-year follow-up period was the primary response variable. The mortality rate in the control group was estimated to be 18% ($p_C=0.18$) and the intervention was believed to have the potential for reducing p_C by 28% ($k=0.28$) yielding $p_I=0.1296$. These estimates of p_C and k were derived from previous studies. Those studies also indicated that the drop-out rate might be as high as 26% over the 3 years; 12% in the first year, an additional 8% in the second year, and an additional 6% in the third year. For the control group, the drop-in rate was estimated to be 7% each year for a total drop-in rate of 21%.

Using these models for adjustment, $p_C^* = 0.1746$ and $p_I^* = 0.1375$. Therefore, instead of δ being 0.0504 $(0.18-0.1296)$, the adjusted δ^* is 0.0371 $(0.1746-0.1375)$. For a two-sided test with $\alpha=0.05$ and $1-\beta=0.90$, the adjusted sample size was 4,020 participants compared to an unadjusted sample size of 2,160 participants. The adjusted sample size almost doubled in this example due to the expected drop-out and drop-in experiences and the recommended policy of keeping participants in the originally assigned study groups. The remarkable increases in sample size because of drop-outs and drop-ins strongly argue for major efforts to keep nonadherence to a minimum during trials.

Sample Size Calculations for Continuous Response Variables

Similar to dichotomous outcomes, we consider two sample size cases for response variables which are continuous [2, 7, 61]. The first case is for two independent samples. The other case is for paired data.

Two Independent Samples

For a clinical trial with continuous response variables, the previous discussion is conceptually relevant, but not directly applicable to actual calculations. "Continuous" variables such as length of hospitalization, blood pressure, spirometric measures, neuropsychological scores, and level of a serum component may be evaluated. Distributions of such measurements frequently can be approximated by a normal distribution. When this is not the case, a transformation of values, such as taking their logarithm, can still make the normality assumption approximately correct.

Suppose the primary response variable, denoted as x, is continuous with N_I and N_C participants randomized to the intervention and control groups, respectively. Assume that the variable x has a normal distribution with mean μ and variance σ^2. The true levels of μ_I and μ_C for the intervention and control groups are not known, but it is assumed that σ^2 is known. (In practice, σ^2 is not known and must be estimated from some data. If the data set used is reasonably large, the estimate of σ^2 can be used in place of the true σ^2. If the estimate for σ^2 is based on a small set of data, it is necessary to be cautious in the interpretation of the sample size calculations.)

The null hypothesis is H_0: $\delta = \mu_C - \mu_I = 0$ and the two-sided alternative hypothesis is H_A: $\delta = \mu_C - \mu_I \neq 0$. If the variance is known, the test statistic is:

$$Z = (\overline{x}_c - \overline{x}_i) / \sigma \sqrt{1/N_C + 1/N_I}$$

This statistic for adequate sample size (e.g., 50 participants per arm) has approximately a standard normal distribution where $\overline{x}_I$ and $\overline{x}_C$ represent mean levels observed in the intervention and control groups, respectively. The hypothesis-testing concepts previously discussed apply to the above statistic. If $Z > Z_\alpha$, then an investigator would reject H_0 at the α level of significance. Using the above test statistic, it can be determined how large a total sample $2N$ would be needed to detect a true difference δ between μ_I and μ_C with power $(1-\beta)$ and significance level α by the formula:

$$2N = \frac{4(Z_\alpha + Z_\beta)^2 \sigma^2}{\delta^2}$$

E*xample*: Suppose an investigator wishes to estimate the sample size necessary to detect a 10 mg/dl difference in cholesterol level in a diet intervention group

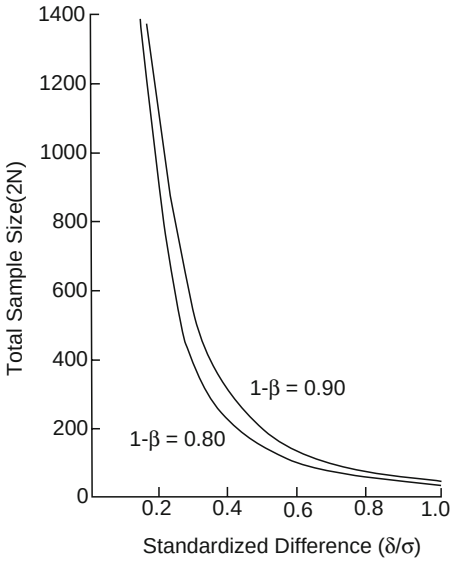


Fig. 8.3 Total sample size ($2N$) required to detect the difference (δ) between control group mean and intervention group mean as a function of the standardized difference (δ/σ) where σ is the common standard deviation, with two-sided significance level of 0.05 and power ($1-\beta$) of 0.80 and 0.90

compared to the control group. The variance from other data is estimated to be $(50\ \text{mg/dl})^2$. For a two-sided 5% significance level, $Z_\alpha = 1.96$ and for 90% power, $Z_\beta = 1.282$. Substituting these values into the above formula, $2N = 4(1.96 + 1.282)^2(50)^2/10^2$ or approximately 1,050 participants. As δ decreases, the value of $2N$ increases, and as σ^2 increases the value of $2N$ increases. This means that the smaller the difference in intervention effect an investigator is interested in detecting and the larger the variance, the larger the study must be. As with the dichotomous case, setting a smaller α and larger $1-\beta$ also increases the sample size. Figure 8.3 shows total sample size $2N$ as a function of δ/σ. As in the example, if $\delta = 10$ and $\sigma = 50$, then $\delta/\sigma = 0.2$ and the sample size $2N$ for $1-\beta = 0.9$ is approximately 1,050.

Paired Data

In some clinical trials, paired outcome data may increase power for detecting differences because individual or within participant variation is reduced. Trial participants may be assessed at baseline and at the end of follow-up. For example, instead of looking at the difference between mean levels in the groups, an investigator interested in mean levels of change might want to test whether diet intervention lowers serum cholesterol from baseline levels when compared to a control. This is essentially the same question that was asked before in the two independent sample

case, but each participant's initial cholesterol level is taken into account. Because of the likelihood of reduced variability, this type of design can lead to a smaller sample size if the question is correctly posed. Assume that Δ_C and Δ_I, represent the true, but unknown levels of change from baseline to some later point in the trial for the control and intervention groups, respectively. Estimates of Δ_C and Δ_I would be $\overline{d}_C = \overline{x}_{C_1} - \overline{x}_{C_2}$ and $\overline{d}_I = \overline{x}_{I_1} - \overline{x}_{I_2}$. These represent the differences in mean levels of the response variable at two points for each group. The investigator tests H_0: $\Delta_C - \Delta_I = 0$ versus H_A: $\Delta_C - \Delta_I = \delta \neq 0$. The variance σ_Δ^2 in this case reflects the variability of the change, from baseline to follow-up, and is assumed here to be the same in the control and intervention arms. This variance is likely to be smaller than the variability at a single measurement. This is the case if the correlation between the first and second measurements is greater than 0.5. Using δ and σ_Δ^2, as defined in this manner, the previous sample size formula for two independent samples and graph are applicable. That is, the total sample size $2N$ can be estimated as

$$2N = \frac{4(Z_\alpha + Z_\beta)^2 \sigma_\Delta^2}{\delta^2}$$

Another way to represent this is

$$2N = \frac{8(Z_\alpha + Z_\beta)^2 (1-\rho)\sigma^2}{\delta^2}$$

where $\sigma_\Delta^2 = 2\sigma^2(1-\rho)$ and σ^2 = the variance of a measurement at a single point in time, the variability is assumed to be the same at either time point (i.e., at baseline and at follow-up), and ρ is the correlation coefficient between the first and second measurement. As indicated, if the correlation coefficient is greater than 0.5, comparing the paired differences will result in a smaller sample size than just comparing the mean values at the time of follow-up.

Example: Assume that an investigator is still interested in detecting a 10 mg/dl difference in cholesterol between the two groups, but that the variance of the change is now (20 mg/dl) [2]. The question being asked in terms of δ is approximately the same because randomization should produce in each group baseline mean levels that are almost equal. The comparison of differences in change is essentially a comparison of the difference in mean levels of cholesterol at the second measurement. Using Fig. 8.3, where $\delta/\sigma_\Delta = 10/20 = 0.5$, the sample size is 170. This impressive reduction in sample size from 1,050 is due to a reduction in the variance from $(50 \text{ mg/dl})^2$to $(20 \text{ mg/dl})^2$.

Another type of pairing occurs in diseases that affect paired organs such as lungs, kidneys, and eyes. In ophthalmology, for example, trials have been conducted where one eye is randomized to receive treatment and the other to receive control therapy [42–45]. Both the analysis and the sample size estimation need to take account of this special kind of stratification. For continuous outcomes, a mean difference in outcome between a treated eye and untreated eye would measure the

treatment effect and could be compared using a paired *t*-test [2, 7], $Z = \overline{d} / S_d \sqrt{1/N}$, where $\overline{d}$ is the average difference in response and S_d is the standard deviation of the differences. The mean difference μ_d is equal to the mean response of the treated or intervention eye, for example, minus the mean response of the control eye, that is $\mu_d = \mu_I - \mu_C$. Under the null hypothesis, μ_d equals δ_d. An estimate of δ_d, $\overline{d}$, can be obtained by taking an estimate of the average differences or by calculating $\overline{x}_I - \overline{x}_C$. The variance of the paired differences σ_d^2 is estimated by S_d^2. Thus, the formula for paired continuous outcomes within an individual is a slight modification of the formula for comparison of means in two independent samples. To compute sample size, N_d, for number of pairs, we compute:

$$N_d = \frac{(Z_\alpha + Z_\beta)^2 \sigma_d^2}{\delta_d^2}$$

As discussed previously, participants in clinical trials do not always fully adhere with the intervention being tested. Some fraction (R_O) of participants on intervention drop-out of the intervention and some other fraction (R_I) drop-in and start following the intervention. If we assume that these participants who drop-out respond as if they had been on control and those who drop-in respond as if they had been on intervention, then the sample size adjustment is the same as for the case of proportions. That is, the adjusted sample size N^* is a function of the drop-out rate, the drop-in rate, and the sample size N for a study with fully compliant participants.

$$N^* = N / (1 - R_O - R_I)^2$$

Therefore, if the drop-out rate were 0.20 and the drop-in 0.05, then the original sample size N must be increased by 16/9 or 1.78; that is, a 78% increase in sample size.

Sample Size for Repeated Measures

The previous section briefly presented the sample size calculation for trials where only two points, say a baseline and a final visit, are used to determine the effect of intervention and these two points are the same for all study participants. Often, a continuous response variable is measured at each follow-up visit. Considering only the first and last values would give one estimate of change but would not take advantage of all the available data. Many models exist for the analysis of repeated measurements and methods for sample size calculation are available for several of these methods [62–69]. In some cases, the response variable may be categorical. We present one of the simpler models for continuous repeated measurements. While other models are beyond the scope of this book, the basic concepts presented are still useful in thinking about how many participants, how many measurements

per individual, and when they should be taken, are needed. In such a case, one possible approach is to assume that the change in response variable is approximately a linear function of time so that the rate of change can be summarized by a slope. This model is fit to each participant's data by the standard least squares method, and the estimated slope is used to summarize the participant's experience. In planning such a study, the investigator must be concerned about the frequency of the measurement and the duration of the observation period. As discussed by Schlesselman [62], the observed measurement x can be expressed as $x = a + bt + \text{error}$, where $a = \text{intercept}$, $b = \text{slope}$, $t = \text{time}$, and error represents the deviation of the observed measurement from a regression line. This error may be due to measurement variability, biological variability, or the nonlinearity of the true underlying relationship. On the average, this error is expected to be equally distributed around 0 and have a variability denoted as $\sigma^2_{(\text{error})}$. Schlesselman assumes that $\sigma^2_{(\text{error})}$ is approximately the same for each participant.

The investigator evaluates intervention effectiveness by comparing the average slope in one group with the average slope in another group. Obviously, participants in a group will not have the same slope, but the slopes will vary around some average value which reflects the effectiveness of the intervention or control. The amount of variability of slopes over participants is denoted as σ^2_b. If D represents the total time duration for each participant and P represents the number of equally spaced measurements, σ can be expressed as

$$\sigma_b^2 = \sigma_B^2 + \left\{ \frac{12(P-1)\sigma^2_{(error)}}{D^2 P(P+1)} \right\}$$

where σ^2_{B} is the component of participant variance in slope not due to measurement error and lack of a linear fit. The sample size required to detect the difference δ between the average rates of change in the two groups is given by

$$2N = \frac{4(Z_\alpha + Z_\beta)^2}{\delta^2} \left[\sigma_B^2 + \frac{12(P-1)\sigma^2_{(error)}}{D^2 P(P+1)} \right]$$

As in the previous formulas, when δ decreases, $2N$ increases. The factor on the right-hand side relates D and P with the variance components σ^2_{B}, and $\sigma^2_{(\text{error})}$. Obviously as σ^2_{B} and $\sigma^2_{(\text{error})}$ increase, the total sample size increases. By increasing P and D, however, the investigator can decrease the contribution made by $\sigma^2_{(\text{error})}$. The exact choices of P and D will depend on how long the investigator can feasibly follow participants, how many times he can afford to have participants visit a clinic, and other factors. By manipulating P and D, an investigator can design a study which will be the most cost effective for his specific situation.

Example: In planning for a trial, it may be assumed that a response variable declines at the rate of 80 units/year in the control group. Suppose a 25% reduction is anticipated in the intervention group. That is, the rate of change in the intervention group would be 60 units/year. Other studies provided an estimate for $\sigma^2_{(\text{error})}$ of

150 units. Also, suppose data from a study of people followed every 3 months for 1 year ($D=1$) and ($P=5$) gave a value for the standard deviation of the slopes, $\sigma_b=200$. The calculated value of σ_B is then 63 units. Thus, for a 5% significance level and 90% power ($Z_\alpha=1.96$ and $Z_\beta=1.282$), the total sample size would be approximately 630 for a 3-year study with four visits per year ($D=3$, $P=13$). Increasing the follow-up time to 4 years, again with four measurements per year, would decrease the variability with a resulting sample size calculation of approximately 510. This reduction in sample size could be used to decide whether or not to plan a 4-year or a 3-year study.

Sample Size Calculations for "Time to Failure"

For many clinical trials, the primary response variable is the occurrence of an event, and thus the proportion of events in each group may be compared. In these cases, the sample size methods described earlier will be appropriate. In other trials, the time to the event may be of special interest. For example, if the time to death or a nonfatal event can be increased, the intervention may be useful even though at some point the proportion of events in each group are similar. Methods for analysis of this type of outcome are generally referred to as life table or survival analysis methods (see Chap. 15). In this situation, other sample size approaches are more appropriate than that described for dichotomous outcomes [70–91]. At the end of this section, we also discuss estimating the number of events required to achieve a desired power.

The basic approach is to compare the survival curves for the groups. A survival curve may be thought of as a graph of the probability of surviving, or not having an event, up to any given point of time. The methods of analysis now widely used are non-parametric; that is, no mathematical model about the shape of the survival curve is assumed. However, for the purpose of estimating sample size, some assumptions are often useful. A common model assumes that the survival curve, $S(t)$, follows an exponential distribution, $S(t)=e^{-\lambda t}=\exp(-\lambda t)$ where λ is called the hazard rate or force of mortality. Using this model, survival curves are totally characterized by λ. Thus, the survival curves from a control and an intervention group can be compared by testing H_0: $\lambda_C=\lambda_I$. An estimate of λ is obtained as the inverse of the mean survival time. If the median survival time, T_M, is known, the hazard rate λ may also be estimated by $-\ln(0.5)/T_M$. Sample size formulations have been considered by several investigators [70–72]. One simple formula is given by

$$2N=\frac{4(Z_\alpha+Z_\beta)^2}{[\ln(\lambda_C/\lambda_I)]^2}$$

where N is the size of the sample in each group and Z_α and Z_β are defined as before. As an example, suppose one assumes that the force of mortality is 0.30 in the control group and expects it to be 0.20 for the intervention being tested; that is, $\lambda_C/\lambda_I=1.5$. If $\alpha=0.05$ (two-sided) and $1-\beta=0.90$, then $N=128$ or $2N=256$.

The corresponding mortality rates for 5 years of follow-up are 0.7769 and 0.6321, respectively. Using the comparison of two proportions, the total sample size would be 412. Thus, the time to failure method may give a more efficient design, requiring a smaller number of participants.

The method just described assumes that all participants will be followed to the event. With few exceptions, clinical trials with a survival outcome are terminated at time T before all participants have had an event. For those still event-free, the time to event is said to be censored at time T. For this situation, Lachin [2] gives the approximate formula:

$$2N = \frac{2(Z_\alpha + Z_\beta)^2[\varphi(\lambda_C) + \varphi(\lambda_C)]}{(\lambda_1 - \lambda_C)^2}$$

where $\varphi(\lambda) = \lambda^2/(1 - e^{-\lambda T})$ and where $\varphi(\lambda_C)$ or $\varphi(\lambda_1)$ are defined by replacing λ with λ_C or λ_1, respectively. If a 5 year study were being planned ($T = 5$) with the same design specifications as above, then the sample size, $2N$ is equal to 376. Thus, the loss of information due to censoring must be compensated for by increasing the sample size. If the participants are to be recruited continually during the 5 years of the trial, the formula given by Lachin is identical but with $\varphi(\lambda) = \lambda^3 T/(\lambda T - 1 + e^{-\lambda T})$. Using the same design assumptions, we obtain $2N = 620$, showing that not having all the participants at the start requires an additional increase in sample size.

More typically, participants are recruited uniformly over a period of time, T_0, with the trial continuing for a total of T years ($T > T_0$). In this situation, the sample size can be estimated as before using

$$\varphi(\lambda) = \frac{\lambda^2}{1 - \left[e^{-\lambda(T - T_0)} - e^{-\lambda T}\right] \Big/ \lambda T_0}$$

Here, the sample size ($2N$) of 466 is between the previous two examples suggesting that it is preferable to get participants recruited as rapidly as possible to get more follow-up or exposure time.

Further models are given by Lachin [2]. A useful series of nomograms has been published [73] for sample size estimates considering factors such as α, $1 - \beta$, the subject recruitment time, the follow-up period, and the ratio of the hazard rates.

One of the methods used for comparing survival curves is the proportional hazards model or the Cox regression model which is discussed briefly in Chap. 15. For this method, sample size estimates have been published [74, 75]. As it turns out, the formula by Schoenfeld for the Cox model [74] is identical to that given above for the simple exponential case, although developed from a different point of view.

All of the above methods assume that the hazard rate remains constant during the course of the trial. This may not be the case. The Beta-Blocker Heart Attack Trial [48] compared 3-year survival in two groups of participants with intervention starting 1–3 weeks after an acute myocardial infarction. The risk of death was high initially, decreased steadily, and then became relatively constant.

For cases where the event rate is relatively small and the clinical trial will have considerable censoring, most of the statistical information will be in the number of events. Thus, the sample size estimates using simple proportions will be quite adequate. In the Beta-Blocker Heart Attack Trial, the 3 year control group event rate was assumed to be 0.18. For the intervention group, the event rate was assumed to be approximately 0.13. In the situation of $\varphi(\lambda) = \lambda^2(1 - e^{-\lambda T})$, a sample size $2N = 2{,}208$ is obtained, before adjustment for estimated nonadherence. In contrast, the unadjusted sample size using simple proportions is 2,160. Again, it should be emphasized that all of these methods are only approximations and the estimates should be viewed as such.

As the previous example indicates, the power of a survival analysis still is a function of the number of events. The expected number of events $E(D)$ is a function of sample size, hazard rate, recruitment rate, and censoring distribution [2, 83]. Specifically, the expected number of events in the control group can be estimated as

$$E(D) = N\lambda_C^2 / \varphi(\lambda_C)$$

where $\varphi(\lambda_C)$ is defined as before, depending on the recruitment and follow-up strategy. If we assume a uniform recruitment over the interval $(0,T_0)$ and follow-up over the interval $(0,T)$, then $E(D)$ can be written using the most general form for $\varphi(\lambda_C)$:

$$E(D) = N\left[1 - \frac{e^{-\lambda(T-T_0)} - e^{-\lambda T}}{\lambda T_0}\right]$$

This estimate of the number of events can be used to predict the number of events at various time points during the trial including the end of follow-up. This prediction can be compared to the observed number of events in the control group to determine if an adjustment needs to be made to the design. That is, if the number of events early in the trial is larger than expected, the trial may be more powerful than designed or may be stopped earlier than the planned T years of follow-up (see Chap. 16). However, more worrisome is when the observed number of events is smaller than what is expected and needed to maintain adequate power. Based on this early information, the design may be modified to attain the necessary number of events by increasing the sample size or expanding recruitment effort within the same period of time, increasing follow-up, or a combination of both.

This method can be illustrated based on a placebo-controlled trial of congestive heart failure [53]. Severe or advanced congestive heart failure has an expected 1 year event rate of 40%, where the events are all-cause mortality and nonfatal myocardial infarction. A new drug was to be tested to reduce the event rate by 25%, using a two-sided 5% significance level and 90% power. If participants are recruited over 1.5 years ($T_0 = 1.5$) during a 2 year study ($T = 2$) and a constant hazard rate is assumed, the total sample size ($2N$) is estimated to be 820 participants with congestive heart failure. The formula $E(D)$ can be used to calculate that approximately 190 events (deaths plus nonfatal myocardial infarctions) must be observed in the control

Table 8.4 Number of expected events (in the control group) at each interim analysis given different event rates in control group

	Number of expected events			
	Calendar time into study			
Yearly event rate in control group (%)	6 months (N=138/group)	1 year (N=275/group)	1.5 years (N=412/group)	2 years (N=412/group)
40	16	60	124	189
35	14	51	108	167
30	12	44	94	146
25	10	36	78	123

group to attain 90% power. If the first year event rate turns out to be less than 40%, fewer events will be observed by 2 years than the required 190. Table 8.4 shows the expected number of control group events at 6 months and 1 year into the trial for annual event rates of 40%, 35%, 30%, and 25%. Two years is also shown to illustrate the projected number of events at the completion of the study. These numbers are obtained by calculating the number of participants enrolled by 6 months (33% of 400) and 1 year (66% of 400) and multiplying by the $\lambda_C^2 / \varphi(\lambda_C)$ term in the equation for $E(D)$. If the assumed annual event rate of 40% is correct, 60 control group events should be observed at 1 year. However, if at 1 year only 44 events are observed, the annual event rate might be closer to 30% (i.e., $\lambda = 0.357$) and some design modification should be considered to assure achieving the desired 190 control group events. One year would be a sensible time to make this decision, based only on control group events since recruitment efforts are still underway. For example, if recruitment efforts could be expanded to 1,220 participants in 1.5 years, then by 2 years of follow-up the 190 events in the placebo group would be observed and the 90% power maintained. If recruitment efforts were to continue for another 6 months at a uniform rate ($T_0 = 2$ years), another 135 participants would be enrolled. In this case, $E(D)$ is $545 \times 0.285 = 155$ events, which would not be sufficient without some additional follow-up. If recruitment and follow-up continued for 27 months (i.e., $T_0 = T = 2.25$), then 605 control group participants would be recruited and $E(D)$ would be 187, yielding the desired power.

Assumptions:

1. Time to event exponentially distributed
2. Uniform entry into the study over 1.5 years
3. Total duration of 2 years

Sample Size for Testing "Equivalency" or Noninferiority of Interventions

In some instances, an effective intervention has already been established and is considered the standard. New interventions under consideration may be preferred because they are less expensive, have fewer side effects, or have less adverse impact

on an individual's general quality of life. This issue is common in the pharmaceutical industry where a product developed by one company may be tested against an established intervention manufactured by another company. Studies of this type are sometimes referred to as trials with positive controls or as noninferiority designs (see Chaps. 3 and 5).

Given that several trials have shown that certain beta-blockers are effective in reducing mortality in post-myocardial infarction participants [48, 92, 93], it is likely that any new beta-blockers developed will be tested against proven agents. The Nocturnal Oxygen Therapy Trial [94] tested whether the daily amount of oxygen administered to chronic obstructive pulmonary disease participants could be reduced from 24 to 12 h without impairing oxygenation. The Intermittent Positive Pressure Breathing [49] trial considered whether a simple and less expensive method for delivering a bronchodilator into the lungs would be as effective as a more expensive device. A breast cancer trial compared the tumor regression rates between subjects receiving the standard, diethylstilbestrol, or the newer agent, tamoxifen [95].

The problem in designing noninferiority trials is that there is no statistical method to demonstrate complete equivalence. That is, it is not possible to show $\delta=0$. Failure to reject the null hypothesis is not a sufficient reason to claim two interventions to be equal but merely that the evidence is inadequate to say they are different [96]. Assuming no difference when using the previously described formulas results in an infinite sample size.

While demonstrating perfect equivalence is an impossible task, one possible approach has been discussed for noninferiority designs [97–99]. The strategy is to specify some value, δ, such that interventions with differences which are less than this might be considered "equally effective" or "noninferior" (see Chap. 5 for discussion of noninferiority designs). Specification of δ may be difficult, but it is a necessary element of the design. The null hypothesis states that $p_C > p_I + \delta$ while the alternative specifies $p_C < p_I + \delta$. The methods developed require that if the two interventions really are equally effective or at least noninferior, the upper $100(1-\alpha)\%$ confidence interval for the intervention difference will not exceed δ with the probability of $1-\beta$. One can alternatively approach this from a hypothesis testing point of view, stating the null hypothesis that the two interventions differ by less than δ.

For studies with a dichotomous response, one might assume the event rate for the two interventions to be equal to p (i.e., $p=p_C=p_I$). This simplifies the previously shown sample size formula to

$$2N = 4p(1-p)(Z_\alpha + Z_\beta)^2 / \delta^2$$

where N, Z_α and Z_β are defined as before. Makuch and Simon [97] recommend for this situation that $\alpha=0.10$ and $\beta=0.20$. However, for many situations, β or Type II error needs to be 0.10 or smaller in order to be sure a new therapy is correctly determined to be equivalent to an older standard. We prefer an $\alpha=0.05$, but this is a matter of judgment and will depend on the situation. (This formula differs slightly from its analogue presented earlier due to the different way the hypothesis is stated.) The formula for continuous variables,

$$2N = \frac{4(Z_\alpha + Z_\beta)^2}{(\delta/\sigma)^2}$$

is identical to the formula for determining sample size discussed earlier. Blackwelder and Chang [99] give graphical methods for computing sample size estimates for studies of equivalency.

Another proposed strategy for comparing a new to a standard drug is to show bioequivalence or similarity in bioavailability. Several authors have discussed this approach [100–103]. If two formulations are within specified limits for a profile of biochemical measurements and one of them has already been proven to be effective, the argument is made that further efficacy trials are not necessary. The sample size estimation for demonstrating bioequivalence poses the same problem as described above and the approach is similar.

As mentioned above and in Chap. 5, specifying δ is a key part of the design and sample size calculations of all equivalency and noninferiority trials. Trials should be sufficiently large, with enough power, to address properly the questions about equivalence or noninferiority that are posed.

Sample Size for Cluster Randomization

So far, sample size estimates have been presented for trials where individuals are randomized. For some prevention trials or health care studies, it may not be possible to randomize individuals. For example, a trial of smoking prevention strategy for teenagers may be implemented most easily by randomizing schools, some schools to be exposed to the new prevention strategy while other schools remain with a standard approach. Individual students are grouped or clustered within each school. As Donner et al. [104] point out, "Since one cannot regard the individuals within such groups as statistically independent, standard sample size formulas underestimate the total number of subjects required for the trial." Several authors [104–107] have suggested incorporating a single inflation factor in the usual sample size calculation to account for the cluster randomization. That is, the sample size per intervention arm N computed by previous formulas will be adjusted to N^* to account for the randomization of N_m clusters, each of size m.

A continuous response is measured for each individual within a cluster of these components. Differences of individuals within a cluster and differences of individuals between clusters contribute to the overall variability of the response. We can separate between - cluster variance σ_b^2 and within - cluster variance σ_w^2. Estimates are denoted by S_b^2 and S_w^2, respectively, and can be estimated by standard analysis of variance. One measure of the relationship of these components is the intra-class correlation coefficient. The intra-class correlation coefficient ρ is $\sigma_b^2/(\sigma_w^2+\sigma_b^2)$ where $0 \le \rho \le 1$. If $\rho = 0$, all clusters respond identically so all of the variability is within a cluster. If $\rho = 1$, all individuals in a cluster respond alike so there is no

variability within a cluster. Estimates of ρ are given by $r = S_b^2 / (S_b^2 + S_w^2)$. Intra-class correlation may range from 0.1 to 0.4 in typical clinical studies. If we computed the sample size calculations assuming no clustering, the sample size per arm would be N participants per treatment arm. Now, instead of randomizing N individuals, we want to randomize N_m clusters each of size m individuals for a total of $N^* = N_m \times m$ participants per treatment arm. The inflation factor [101] is $[1+(m-1)r]$ so that

$$N^* = N_m \times m = N\left[1+(m-1)\rho\right]$$

Note that the inflation factor is a function of both cluster size m and intra-class correlation. If the intra-cluster correlation ($\rho=0$), then each individual in one cluster responds like any individual in another cluster, and the inflation factor is unity ($N^*=N$). That is, no penalty is paid for the convenience of cluster randomization. At the other extreme, if all individuals in a cluster respond the same ($\rho=1$), there is no added information within each cluster, so only one individual per cluster is needed, and the inflation factor is m. That is, our adjusted sample $N^*=N\times m$ and we pay a severe price for this type of cluster randomization. However, it is unlikely that ρ is either 0 or 1, but as indicated, is more likely to be in the range of 0.1–0.4 in clinical studies.

Example: Donner et al. [104] provide an example for a trial randomizing households to a sodium reducing diet in order to reduce blood pressure. Previous studies estimated the intra-class correlation coefficient to be 0.2; that is $\hat{\rho} = r = S_b^2 / (S_b^2 + S_w^2) = 0.2$. The average household size was estimated at 3.5 ($m=3.5$). The sample size per arm N must be adjusted by $1+(m-1)\rho = 1+(3.5-1)(0.2) = 1.5$. Thus, the normal sample size must be inflated by 50% to account for this randomization indicating a small between cluster variability. If $\rho=0.1$, then the factor is $1+(3.5-1)(0.1)$ or 1.25. If $\rho=0.4$, indicating a larger between cluster component of variability, the inflation factor is 2.0 or a doubling.

For binomial responses, a similar expression for adjusting the standard sample size can be developed. In this setting, a measure of the degree of within cluster dependency or concordancy rate in participant responses is used in place of the intra-class correlation. The commonly used measure is the kappa coefficient, denoted κ, and may be thought of as an intra-class correlation coefficient for binomial responses, analogous to ρ for continuous responses. A concordant cluster with $\kappa=1$ is one where all responses within a cluster are identical, all successes or failures, in which a cluster contributes no more than a single individual. A simple estimate for κ is provided [104].

$$\kappa = \frac{p^*\left[p_C^m + (1-p_C)^m\right]}{1-\left[p_C^m + (1-p_C)^m\right]}$$

Here p^* is the proportion of the control group with concordant clusters, and p_C is the underlying success rate in the control group. The authors then show that the inflation factor is $[1+(m-1)\kappa]$, or that the regular sample size per treatment arm N must be multiplied by this factor to attain the adjust sample size N^*.

$$N^* = N\left[1 + (m-1)\kappa\right]$$

Example: Donner et al. [104] continues the sodium diet example where couples ($m=2$) are randomized to either a low sodium or a normal diet. The outcome is the hypertension rate. Other data suggest the concordancy of hypertension status among married couples is 0.85 ($p^*=0.85$). The control group hypertension rate is 0.15 ($p_C=0.15$). In this case, $\kappa=0.41$, so that the inflation factor is $1+(2-1)(0.41)=1.41$; that is, the regular sample size must be inflated by 41% to adjust for the couples being the randomization unit. If there is perfect control group concordance, $p^*=1$ and $\kappa=1$, in which case, $N^*=2N$.

Cornfield proposed another adjustment procedure [107]. Consider a trial where m clusters will be randomized, each cluster of size c_i ($i=1, 2, \ldots, m$) and each having a different success rate of p_i ($i=1, 2, \ldots, m$). Define the average cluster size $\overline{c} = \Sigma c_i / m$ and $\overline{p} = \Sigma c_i p_i \, \Sigma c_i$ as the overall success rate weighted by cluster size. The variance of the overall success rate $\sigma_p^2 = \Sigma c_i (p_i - \overline{p})^2 / m\overline{c}^2$. In this setting, the efficiency of simple randomization to cluster randomization is $E = \overline{p}(1-\overline{p})^2 \overline{c}\sigma_p^2$. The inflation factor (IF) for this design is $\mathrm{IF} = 1/E = \overline{c}\sigma_p^2 / (1-\overline{p})$. Note that if the response rate varies across clusters, the normal sample size must be increased.

While cluster randomization may be logistically required, the process of making the cluster the randomization unit has serious sample size implications. It would be unwise to ignore this consequence in the design phase. As shown, the sample size adjustments can easily be factors of 1.5 or higher. For clusters which are schools or cities, the intra-class correlation is likely to be quite small. However, the cluster size is multiplied by the intra-class correlation so that the impact might still be non-trivial. Not making this adjustment would substantially reduce the study power if the analyzes were done properly, taking into account the cluster effect. Ignoring the cluster effect in the analysis would be viewed critically in most cases and is not recommended.

Estimating Sample Size Parameters

As shown in the methods presented, sample size estimation is quite dependent upon assumptions made about variability of the response, level of response in the control group, and the difference anticipated or judged to be clinically relevant [10, 108–113]. Obtaining reliable estimates of variability or levels of response can be challenging since the information is often based on very small studies or studies not exactly relevant to the trial being designed. Sometimes, pilot or feasibility studies may be conducted to obtain these data. In such cases, the term external pilot has been used [113].

In some cases, the information may not exist prior to starting the trial, as was the case for early trials in AIDS; that is, no incidence rates were available in an evolving epidemic. Even in cases where data are available, other factors affect the variability

or level of response observed in a trial. Typically, the variability observed in the planned trial is larger than expected or the level of response is lower than assumed. Numerous examples of this experience exist [108]. One is provided by the Physicians' Health Study [114]. In this trial, 22,000 US male physicians were randomized into a 2×2 factorial design. One factor was aspirin versus placebo in reducing cardiovascular mortality. The other factor was beta-carotene versus placebo for reducing cancer incidence. The aspirin portion of the trial was terminated early in part due to a substantially lower mortality rate than expected. In the design, the cardiovascular mortality rate was assumed to be approximately 50% of the US age-adjusted rate in men. However, after 5 years of follow-up, the rate was approximately 10% of the US rate in men. This substantial difference reduced the power of the trial dramatically. In order to compensate for the extremely low event rate, the trial would have had to be extended another 10 years to get the necessary number of events [114]. One can only speculate about reasons for low event rates, but screening of potential participants prior to the entry almost certainly played a part. That is, screens had to complete a run-in period and be able to tolerate aspirin. Those at risk for other competing events were also excluded. This type of effect is referred to as a screening effect. Physicians who began to develop cardiovascular signs may have obtained care earlier than non-physicians. In general, volunteers for trials tend to be healthier than the general population, a phenomenon often referred to as the healthy volunteer effect.

Another approach to obtaining estimates for ultimate sample size determination is to design so-called internal pilot studies [113]. In this approach, a small study is initiated based on the best available information. A general sample target for the full study may be proposed, but the goal of the pilot is to refine that sample size estimate based on screening and healthy volunteer effects. The pilot study uses a protocol very close if not identical to the protocol for the full study, and thus parameter estimates will reflect those effects. If the protocol for the pilot and the main study are essentially identical, then the small pilot can become an internal pilot. That is, the data from the internal pilot become part of the data for the overall study. This approach was used successfully in the Diabetes Control and Complications Trial [115]. If data from the internal pilot are used only to refine estimates of variability or control group response rates, and not changes in treatment effect, then the impact of this two step approach on the significance level is negligible. However, the benefit is that this design will more likely have the desired power than if data from external pilots and other sources are relied on exclusively [112]. It must be emphasized that pilot studies, either external or internal, should not be viewed as providing reliable estimates of the intervention effect [113]. Because power is too small in pilot studies to be sure that no effect exists, small or no differences may erroneously be viewed as reason not to pursue the question. A positive trend may also be viewed as evidence that a large study is not necessary, or that clinical equipoise no longer exists.

Our experience indicates that both external and internal pilot studies are quite helpful. Internal pilot studies should be used if at all possible in prevention trials, when screening and healthy volunteer effects seem to cause major design problems.

Design modifications based on an internal pilot are more prudent than allowing an inadequate sample size to create yield misleading results.

One approach is to specify the number of events needed for a desired power level. Obtaining the specified number of events requires estimating the number of individuals to be followed for a period of time. How many participants are involved and for how long they are followed can be adjusted during the early part of the trial, or during an internal pilot study, but the target number of events remains unchanged. This is also discussed in more detail in Chap. 16.

Another approach is to use adaptive designs which modify the sample size based on an emerging trend, referred to as trend adaptive designs (see Chaps. 5 and 16). Here the sample size may be adjusted for an updated estimate of the treatment effect, δ, using the methods described in this chapter. However, an adjustment must then be made at the analysis stage which may require a substantially larger critical value than the standard one in order to maintain a prespecified α level.

Multiple Response Variables

We have stressed the advantages of having a single primary question and a single primary response variable, but clinical trials occasionally have more than one of each. More than one question may be asked because investigators cannot agree about which outcome is most important. As an example, one clinical trial involving two schedules of oxygen administration to participants with chronic obstructive pulmonary disease had three major questions in addition to comparing the mortality rate [94]. Measures of pulmonary function, neuro-psychological status, and quality of life were evaluated. For the participants, all three were important.

Sometimes more than one primary response variable is used to assess a single primary question. This may reflect uncertainty as to how the investigator can answer the question. A clinical trial involving participants with pulmonary embolism [116] employed three methods of determining a drug's ability to resolve emboli. They were: lung scanning, arteriography, and hemodynamic studies. Another trial involved the use of drugs to limit myocardial infarct size [117]. Precordial electrocardiogram mapping, radionuclide studies, and enzyme levels were all used to evaluate the effectiveness of the drugs.

Computing a sample size for such clinical trials is not easy. One could attempt to define a single model for the multidimensional response and use one of the previously discussed formulas. Such a method would require several assumptions about the model and its parameters and might require information about correlations between different measurements. Such information is rarely available. A more reasonable procedure would be to compute sample sizes for each individual response variable. If the results give about the same sample size for all variables, then the issue is resolved. However, more commonly, a range of sample sizes will be obtained. The most conservative strategy would be to use the largest sample size computed. The other response variables would then have even greater power to detect the hoped-for

reductions or differences (since they required smaller sample sizes). Unfortunately, this approach is the most expensive and difficult to undertake. Of course, one could also choose the smallest sample size of those computed. That would probably not be desirable because the other response variables would have less power than usually required, or only larger differences than expected would be detectable. It is possible to select a middle range sample size, but there is no assurance that this will be appropriate. An alternative approach is to look at the difference between the largest and smallest sample sizes. If this difference is very large, the assumptions that went into the calculations should be re-examined and an effort should be made to resolve the difference.

As discussed in Chap. 17, when multiple comparisons are made, the chance of finding a significant difference in one of the comparisons (when, in fact, no real differences exist between the groups) is greater than the stated significance level. In order to maintain an appropriate significance level α for the entire study, the significance level required for each test to reject H_0 should be adjusted [22]. The significance level required for rejection (α') in a single test can be approximated by α/k where k is the number of multiple response variables. For several response variables, this can make α' fairly small (e.g., $k=5$ implies $\alpha'=0.01$ for each of k response variables with an overall $\alpha=0.05$). If the correlation between response variables is known, then the adjustment can be made more precisely [118]. In all cases, the sample size would be much larger than if the use of multiple response variables were ignored, so most studies have not strictly adhered to this solution of modifying the significance level. Some investigators, however, have attempted to be conservative in the analysis of results [119]. There is a reasonable limit as to how much α' can be decreased in order to give protection against false rejection of the null hypothesis. Some investigators have chosen $\alpha'=0.01$ regardless of the number of tests. In the end, there are no easy solutions. A somewhat conservative value of α' needs to be set, and the investigators need to be aware of the multiple testing problem during the analysis.

References

1. Freiman JA, Chalmers TC, Smith H, Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. *N Engl J Med* 1978;299:690–694.
2. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. *Control Clin Trials* 1981;2:93–113.
3. Brown BW, Jr. Statistical controversies in the design of clinical trials – some personal views. *Control Clin Trials* 1980;1:13–27.
4. Altman DG. Statistics and ethics in medical research: III. How large a sample? *Br Med J* 1980;281:1336–1338.
5. Gore SM. Assessing clinical trials – trial size. *Br Med J* 1981;282:1687–1689.
6. Day SJ, Graham DF. Sample size estimation for comparing two or more treatment groups in clinical trials. *Stat Med* 1991;10:33–43.

7. Donner A. Approaches to sample size estimation in the design of clinical trials – a review. *Stat Med* 1984;3:199–214.
8. Phillips AN, Pocock SJ. Sample size requirements for prospective studies, with examples for coronary heart disease. *J Clin Epidemiol* 1989;42:639–648.
9. Steiner DL. Sample size and power in psychiatric research. *Can J Psychiatry* 1990;35:616–620.
10. Whitehead J. Sample sizes for phase II and phase III clinical trials: An integrated approach. *Stat Med* 1986;5:459–464.
11. Schlesselman JJ. Planning a longitudinal study: I. Sample size determination. *J Chronic Dis* 1973;26:553–560.
12. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions* (3rd ed.). New York: John Wiley and Sons, 2003.
13. Snedecor GW, Cochran WG. *Statistical Methods* (8th ed.). Ames: Iowa State University Press, 1989.
14. Brown BW, Hollander M. *Statistics – A Biomedical Introduction*. New York: John Wiley and Sons, 1977.
15. Remington RD, Schork MA. *Statistics with Applications to the Biological and Health Sciences*. Englewood Cliffs, NJ: Prentice-Hall, 1970.
16. Dixon WJ, Massey FJ, Jr. *Introduction to Statistical Analysis* (3rd ed.). New York: McGraw-Hill, 1969.
17. Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research* (4th ed.). Malden, MA: Blackwell Publishing, 2002.
18. Woolson RF. *Statistical Methods for the Analysis of Biomedical Data*. New York: John Wiley and Sons, 1987.
19. Fisher L, Van Belle G. *Biostatistics – A Methodology for the Health Sciences*. New York: John Wiley and Sons, 1993.
20. Canner PL, Klimt CR. Experimental design features. *Control Clin Trials* 1983;4:313–332.
21. Dunnett CW. Multiple comparison procedures for comparing several treatments with a control. *J Am Stat Assoc* 1955;50:1096–1121.
22. Costigan T. Bonferroni inequalities and intervals. In Armitage P, Colton T (eds.). *Encyclopedia of Biostatistic*s (2nd ed.). New York: John Wiley & Sons, 2007.
23. Brittain E, Schlesselman JJ. Optimal allocation for the comparison of proportions. *Biometrics* 1982;38:1003–1009.
24. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N Engl J Med* 1989;321:406–412.
25. Rothman KJ. A show of confidence. *N Engl J Med* 1978;299:1362–1363.
26. McHugh RB, Le CT. Confidence estimation and the size of a clinical trial. *Control Clin Trials* 1984;5:157–163.
27. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. *J R Stat Soc Ser A* 1969;132:235–244.
28. Gail M, Gart JJ. The determination of sample sizes for use with the exact conditional test in 2×2 comparative trials. *Biometrics* 1973;29:441–448.
29. Gail M. The determination of sample sizes for trials involving several independent 2×2 tables. *J Chronic Dis* 1973;26:669–673.
30. Haseman JK. Exact sample sizes for use with the Fisher–Irwin Test for 2×2 tables. *Biometrics* 1978;34:106–109.
31. Feigl P. A graphical aid for determining sample size when comparing two independent proportions. *Biometrics* 1978;34:111–122.
32. Casagrande JT, Pike MC. An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics* 1978;34:483–486.
33. Ury HK, Fleiss JL. On approximate sample sizes for comparing two independent proportions with the use of Yates' correction. *Biometrics* 1980;36:347–351.

34. Fleiss JL, Tytun A, Ury HK. A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics* 1980;36:343–346.
35. Wacholder S, Weinberg CR. Paired versus two-sample design for a clinical trial of treatments with dichotomous outcome: Power considerations. *Biometrics* 1982;38:801–812.
36. Day SJ. Optimal placebo response rates for comparing two binomial proportions. *Stat Med* 1988;7:1187–1194.
37. Fu YX, Arnold J. A table of exact sample sizes for use with Fisher's exact test for 2×2 tables. *Biometrics* 1992;48:1103–1112.
38. Lachenbruch PA. A note on sample size computation for testing interactions. *Stat Med* 1988;7:467–469.
39. McMahon RP, Proschan M, Geller NL, et al. Sample size calculation for clinical trials in which entry criteria and outcomes are counts of events. *Stat Med* 1994;13:859–870.
40. Bristol DR. Sample sizes for constructing confidence intervals and testing hypotheses. *Stat Med* 1989;8:803–811.
41. Connor RJ. Sample size for testing differences in proportions for the paired-sample design. *Biometrics* 1987;43:207–211.
42. Donner A. Statistical methods in ophthalmology: An adjusted chi-square approach. *Biometrics* 1989;45:605–611.
43. Gauderman WJ, Barlow WE. Sample size calculations for ophthalmologic studies. *Arch Ophthalmol* 1992;110:690–692.
44. Rosner B. Statistical methods in ophthalmology: An adjustment for the intraclass correlation between eyes. *Biometrics* 1982;38:105–114.
45. Rosner B, Milton RC. Significance testing for correlated binary outcome data. *Biometrics* 1988;44:505–512.
46. Coronary Drug Project Research Group. The Coronary Drug Project. Design, methods, and baseline results. *Circulation* 1973;47:I1–I50.
47. Aspirin Myocardial Infarction Study Research Group. A randomized, controlled trial of aspirin in persons recovered from myocardial infarction. *JAMA* 1980;243:661–669.
48. Beta-Blocker Heart Attack Trial Research Group. A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. *JAMA* 1982;247:1707–1714.
49. The Intermittent Positive Pressure Breathing Trial Group. Intermittent positive pressure breathing therapy of chronic obstructive pulmonary disease – a clinical trial. *Ann Intern Med* 1983;99:612–620.
50. CASS Principal Investigators and Their Associates. Coronary Artery Surgery Study (CASS): A randomized trial of coronary artery bypass surgery. Survival data. *Circulation* 1983;68:939–950.
51. Hypertension Detection and Follow-up Program Cooperative Group. Five-year findings of the Hypertension Detection and Follow-up Program. I. Reduction in mortality of persons with high blood pressure, including mild hypertension. *JAMA* 1979;242:2562–2571.
52. Multiple Risk Factor Intervention Trial Research Group. Multiple risk factor intervention trial. Risk factor changes and mortality results. *JAMA* 1982;248:1465–1477.
53. Packer M, Carver JR, Rodeheffer RJ, et al. for the PROMISE Study Research Group. Effect of oral milrinone on mortality in severe chronic heart failure. *N Engl J Med* 1991;325:1468–1475.
54. Halperin M, Rogot E, Gurian J, Ederer F. Sample sizes for medical trials with special reference to long-term therapy. *J Chronic Dis* 1968;21:13–24.
55. Schork MA, Remington RD. The determination of sample size in treatment–control comparisons for chronic disease studies in which drop-out or non-adherence is a problem. *J Chronic Dis* 1967;20:233–239.
56. Wu M, Fisher M, DeMets D. Sample sizes for long-term medical trial with time-dependent dropout and event rates. *Control Clin Trials* 1980;1:111–124.
57. Barlow W, Azen S. for the Silicone Study Group. The effect of therapeutic treatment crossovers on the power of clinical trials. *Control Clin Trials* 1990;11:314–326.
58. Lakatos E. Sample size determination in clinical trials with time-dependent rates of losses and noncompliance. *Control Clin Trials* 1986;7:189–199.

59. Lavori P. Statistical issues: Sample size and drop out. In Benkert O, Maier W, Rickels K (eds.). *Methodology of the Evaluation of Psychotropic Drugs.* Berlin: Springer-Verlag, 1990.
60. Newcombe RG. Explanatory and pragmatic estimates of the treatment effect when deviations from allocated treatment occur. *Stat Med* 1988;7:1179–1186.
61. Pentico DW. On the determination and use of optimal sample sizes for estimating the difference in means. *Am Stat* 1981;35:40–42.
62. Schlesselman JJ. Planning a longitudinal study: II. Frequency of measurement and study duration. *J Chronic Dis* 1973;26:561–570.
63. Dawson JD, Lagakos SW. Size and power of two-sample tests of repeated measures data. *Biometrics* 1993;49:1022–1032.
64. Kirby AJ, Galai N, Muñoz A. Sample size estimation using repeated measurements on biomarkers as outcomes. *Control Clin Trials* 1994;15:165–172.
65. Laird NM, Wang F. Estimating rates of change in randomized clinical trials. *Control Clin Trials* 1990;11:405–419.
66. Lipsitz SR, Fitzmaurice GM. Sample size for repeated measures studies with binary responses. *Stat Med* 1994;13:1233–1239.
67. Nam J. A simple approximation for calculating sample sizes for detecting linear trend in proportions. *Biometrics* 1987;43:701–705.
68. Overall JE, Doyle SR. Estimating sample sizes for repeated measurement designs. *Control Clin Trials* 1994;15:100–123.
69. Rochon J. Sample size calculations for two-group repeated-measures experiments. *Biometrics* 1991;47:1383–1398.
70. Pasternack BS, Gilbert HS. Planning the duration of long-term survival time studies designed for accrual by cohorts. *J Chronic Dis* 1971;24:681–700.
71. Pasternack BS. Sample sizes for clinical trials designed for patient accrual by cohorts. *J Chronic Dis* 1972;25:673–681.
72. George SL, Desu MM. Planning the size and duration of a clinical trial studying the time to some critical event. *J Chronic Dis* 1974;27:15–24.
73. Schoenfeld DA, Richter JR. Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics* 1982;38:163–170.
74. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics* 1983;39:499–503.
75. Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. *Stat Med* 1982;1:121–129.
76. Akazawa K, Nakamura T, Moriguchi S. Simulation program for estimating statistical power of Cox's proportional hazards model assuming no specific distribution for the survival time. *Comput Methods Programs Biomed* 1991;35:203–212.
77. Cantor AB. Power estimation for rank tests using censored data: Conditional and unconditional. *Control Clin Trials* 1991;12:462–473.
78. Emerich LJ. Required duration and power determinations for historically controlled studies of survival times. *Stat Med* 1989;8:153–160.
79. Gail MH. Applicability of sample size calculations based on a comparison of proportions for use with the logrank test. *Control Clin Trials* 1985;6:112–119.
80. Gross AJ, Hunt HH, Cantor AB, Clark BC. Sample size determination in clinical trials with an emphasis on exponentially distributed responses. *Biometrics* 1987;43:875–883.
81. Halperin M, Johnson NJ. Design and sensitivity evaluation of follow-up studies for risk factor assessment. *Biometrics* 1981;37:805–810.
82. Hsieh FY. Sample size tables for logistic regression. *Stat Med* 1989;8:795–802.
83. Lachin JM, Foulkes MA. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics* 1986;42:507–519. (Correction: 42:1009, 1986).
84. Lakatos E. Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics* 1988;44:229–241.

85. Lui K-J. Sample size determination under an exponential model in the presence of a confounder and type I censoring. *Control Clin Trials* 1992;13:446–458.
86. Morgan TM. Nonparametric estimation of duration of accrual and total study length for clinical trials. *Biometrics* 1987;43:903–912.
87. Palta M, Amini SB. Consideration of covariates and stratification in sample size determination for survival time studies. *J Chronic Dis* 1985;38:801–809.
88. Rubenstein LV, Gail MH, Santner TJ. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *J Chronic Dis* 1981;34:469–479.
89. Taulbee JD, Symons MJ. Sample size and duration for cohort studies of survival time with covariables. *Biometrics* 1983;39:351–360.
90. Wu MC. Sample size for comparison of changes in the presence of right censoring caused by death, withdrawal, and staggered entry. *Control Clin Trials* 1988;9:32–46.
91. Zhen B, Murphy JR. Sample size determination for an exponential survival model with an unrestricted covariate. *Stat Med* 1994;13:391–397.
92. Hjaimarson A, Herlitz J, Malek I, et al. Effect on mortality of metoprolol in acute myocardial infarction: A double-blind randomized trial. *Lancet* 1981;ii:823–827.
93. The Norwegian Multicenter Study Group. Timolol-induced reduction in mortality and reinfarction in patients surviving acute myocardial infarction. *N Engl J Med* 1981;304:801–807.
94. Nocturnal Oxygen Therapy Trial Group. Continuous or nocturnal oxygen therapy in hypoxemic chronic obstructive lung disease: A clinical trial. *Ann Intern Med* 1980;93:391–398.
95. Ingle JN, Ahmann DL, Green SJ, et al. Randomized clinical trial of diethylstilbestrol versus tamoxifen in postmenopausal women with advanced breast cancer. *N Engl J Med* 1981;304:16–21.
96. Spriet A, Beiler D. When can "non significantly different" treatments be considered as "equivalent"? (Letter to the editors). *Br J Clin Pharmacol Ther* 1979;7:623–624.
97. Makuch R, Simon R. Sample size requirements for evaluating a conservative therapy. *Cancer Treat Rep* 1978;62:1037–1040.
98. Blackwelder WC. "Proving the null hypothesis" in clinical trials. *Control Clin Trials* 1982;3:345–353.
99. Blackwelder WC, Chang MA. Sample size graphs for "proving the null hypothesis". *Control Clin Trials* 1984;5:97–105.
100. Anderson S, Hauck WW. A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Commun Stat Theory Methods* 1983;12:2663–2692.
101. Dunnett CW, Gent M. Significance testing to establish equivalence between treatments, with special reference to data in the form of 2×2 tables. *Biometrics* 1977;33:593–602.
102. Westlake WJ. Statistical aspects of comparative bioavailability trials. *Biometrics* 1979;35:273–280.
103. Kirkwood TBL, Westlake WJ. Response to "Bioequivalence testing-a need to rethink". *Biometrics* 1981;37:589–594.
104. Donner A, Birkett N, Buck C. Randomization by cluster. Sample size requirements and analysis. *Am J Epidemiol* 1981;114:906–914.
105. Hsieh FY. Sample size formulae for intervention studies with the cluster as unit of randomization. *Stat Med* 1988;7:1195–1201.
106. Lee EW, Dubin, N. Estimation and sample size considerations for clustered binary responses. *Stat Med* 1994;13:1241–1252.
107. Cornfield J. Randomization by group: A formal analysis. *Am J Epidemiol* 1978;108:100–102.
108. Church TR, Ederer F, Mandel JS, et al. Estimating the duration of ongoing prevention trials. *Am J Epidemiol* 1993;137:797–810.
109. Ederer F, Church TR, Mandel JS. Sample sizes for prevention trials have been too small. *Am J Epidemiol* 1993;137:787–796.
110. Neaton JD, Bartsch GE. Impact of measurement error and temporal variability on the estimation of event probabilities for risk factor intervention trials. *Stat Med* 1992;11:1719–1729.

111. Patterson BH. The impact of screening and eliminating preexisting cases on sample size requirements for cancer prevention trials. *Control Clin Trials* 1987;8:87–95.
112. Shih WJ. Sample size reestimation in clinical trials. In Peace KE (ed.). *Biopharmaceutical Sequential Statistical Applications*. New York: Marcel Dekker, 1992, pp. 285–301.
113. Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. *Stat Med* 1990;9:65–72.
114. Steering Committee of the Physicians' Health Study Research Group. Final report on the aspirin component of the ongoing Physicians' Health Study. *N Engl J Med* 1989;321:129–135.
115. The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med* 1993;329:977–986.
116. Urokinase Pulmonary Embolism Trial Study Group. Urokinase–streptokinase embolism trial. Phase 2 results. *JAMA* 1974;229:1606–1613.
117. Roberts R, Croft C, Gold HK, et al. Effect of propranolol on myocardial-infarct size in a randomized blinded multicenter trial. *N Engl J Med* 1984;311:218–225.
118. Miller RG. *Simultaneous Statistical Inference*. New York: McGrawHill, 1966.
119. The Coronary Drug Project Research Group. Clofibrate and niacin in coronary heart disease. *JAMA* 1975;231:360–381.

Chapter 9
Baseline Assessment

In clinical trials, baseline refers to the status of a participant before the start of intervention. Baseline data may be measured by interview, questionnaire, physical examination, laboratory tests, and procedures. Measurement need not be only numerical in nature. It can also mean classification of study participants into categories based on factors such as absence or presence of some trait or condition.

As discussed in Chap. 4, baseline data describe the people studied, enabling the scientific community to compare the trial results with those of other studies. Different results from different studies may be attributed to seemingly minor differences in the study participants. These differences, in turn, can lead to the creation of new hypotheses, which may be tested.

Valid inferences about benefits and harmful effects of therapy depend on the kinds of people enrolled, as well as on study design. Complete reporting of baseline data allows clinicians to evaluate a new therapy's chance of success or failure in their patients. On the other hand, judgment should be used in determining the factors to be measured at baseline. Evaluating factors that are unlikely to be pertinent to the trial not only wastes money and time but may also reduce participant cooperation. This chapter is concerned with the uses of baseline data, what constitutes a true baseline measurement, and assessment of baseline comparability.

Fundamental Point

Relevant baseline data should be measured in all study participants before the start of intervention.

Uses of Baseline Data

Although baseline data may be used to determine the eligibility of participants, it is assumed that any participants who are found unable to meet entrance criteria have been excluded from the study before assignment to either intervention or control.

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_9, © Springer Science+Business Media, LLC 2010

The characteristics of people not enrolled are of interest when attempting to generalize the trial results (Chap. 4). For the discussion in this chapter, however, only data from enrolled participants are considered.

The amount of data collected at baseline depends on the nature of the trial and the purpose for which the data will be used. As mentioned elsewhere, some trials have simple protocols and collect limited amounts of data. If such trials are large, it is reasonable to expect that good balance between groups will be achieved. Because the goals of these trials are restricted to answering the primary question and one or two secondary questions, the other uses for baseline data are unnecessary. The investigators do not intend to perform stratification and special subgroup analyses or to conduct natural history studies. The simple design of such studies means that detailed documentation of baseline variables is omitted and only a few key demographic and medical variables are ascertained.

Baseline Comparability

Baseline data allow people to evaluate whether the study groups were comparable before intervention was started. The assessment of comparability typically includes pertinent demographic and socioeconomic characteristics, risk or prognostic factors, medications, and medical history. This assessment is necessary in both randomized and nonrandomized trials. In assessment of comparability in any trial, the investigator can only look at factors about which she is aware. Obviously, those which are unknown cannot be compared. The baseline characteristics of each group should be presented in the main results paper of every randomized trial. Special attention should be given to factors that may influence any benefit of the study intervention and those that may predict adverse events. Full attention to baseline comparability is not always given. In a review of 206 surgical trials, only 73% reported baseline data [1]. Moreover, more than one-quarter of those trials reported fewer than five baseline factors. Altman and Doré, in a review of 80 published randomized trials, noted considerable variation in the quality of the reporting of baseline characteristics [2]. Half of those reporting continuous covariates did not use appropriate measures of variability.

While randomization on the average produces balance between comparison groups, it does not guarantee balance in any specific trial or for any specific baseline measure. However, Lachin [3] argues that the likelihood of baseline imbalance is small if the total sample size is 200 or more. Clearly, imbalances are more common in smaller trials, but they only matter if they modify the trial outcome. A placebo-controlled, double-blind trial in 39 participants with mucopolysaccharidosis type IV reported that the intervention significantly improved endurance [4]. However, the 12-min walk test showed the distance walked at baseline to be 227 m in the intervention group and 381 m in the placebo group. A double-blind placebo-controlled trial in 341 participants with Alzheimer's disease evaluated three active treatments – vitamin E, a selective monoamine oxidase inhibitor and their combination [5].

The Mini-Mental State Examination (MMSE) score, a variable highly predictive of the primary outcome, was significantly higher in the placebo group at baseline than in the other groups. In unadjusted analyses, there were no differences among the groups. After adjustment for the baseline difference in MMSE, all actively treated groups did better than placebo by slowing the progression of disease. Imbalances may even be true in large studies. In the Aspirin Myocardial Infarction Study [6], which had over 4,500 participants, the aspirin group was at slightly higher risk than the placebo group when baseline characteristics were examined.

Stratification

If there is concern that one or two key prognostic factors may not "balance out" during randomization, thus yielding imbalanced groups at baseline, the investigator may stratify on the basis of these factors. Stratification can be done at the time of randomization or during analysis. Chapters 5 and 17 review the advantages and disadvantages of stratified randomization and stratified analysis. The point here is that, in order to stratify at either time, the relevant characteristics of the participants at baseline must be known. For nonrandomized trials, these factors must also be measured in order to select properly the control group and analyze results by strata.

Subgrouping

Often, investigators are interested not only in the response to intervention in the total study group, but also in the response in one or more subgroups. Particularly, in studies in which an overall intervention effect is present, analysis of results by appropriate subgroup may help to identify the specific population most likely to benefit from, or be harmed by, the intervention. Subgrouping may also help to elucidate the mechanism of action of the intervention. Definition of such subgroups should rely only on baseline data, not data measured after initiation of intervention (except for factors such as age or sex which cannot be altered by the intervention). An example of establishing subgroups is the Canadian Cooperative Study Group trial of aspirin and sulfinpyrazone in people with cerebral or retinal ischemic attacks [7]. After noting an overall benefit from aspirin in reducing continued ischemic attacks or stroke, the authors observed that the benefit was restricted to men. Any conclusions drawn from subgroup hypotheses not explicitly stated in the protocol, however, should be given much less credibility than those from hypotheses stated a priori. Retrospective subgroup analyses should serve primarily to generate new hypotheses for subsequent testing (Chap. 17). In approving aspirin for the indication of transient ischemic attacks in men, the U.S. Food and Drug Administration relied on the Canadian Cooperative Study Group. A subsequent meta-analysis of platelet active drug trials in the secondary prevention of cardiovascular

disease concluded that the effect is similar in men and women [8]. However, a later placebo-controlled primary prevention trial of low-dose aspirin (100 mg on alternate days) in women reported a favorable aspirin effect on the risk of stroke, but no reduction in risks of myocardial infarction and cardiovascular death [9].

One of the large active-control trials of rosiglitazone in people with type 2 diabetes reported a surprising increase in the risk of fractures compared to metformin or glibenclamide, a risk that was limited to women [10]. This posthoc observation was replicated in a subsequent trial of pioglitazone, which showed a similar gender-specific increase compared to placebo [11]. A meta-analysis confirmed that this class of hypoglycemic agents doubles the risk of fractures in women without any increase in men [12].

In their review of 50 clinical trial reports from four major medical journals, Assmann et al. [13] noted deficiencies in the presentation of subgroup findings. The median number of subgroup analyses was four; the largest number was 24. More than half of the subgroup reports failed to use statistical tests for interaction. Such tests are critical, since they directly determine whether an observed treatment difference in an outcome depends on the participant's subgroup. Reliance on *p*-values for treatment difference in each separate subgroup is not appropriate. Many articles placed too much emphasis on subgroup findings and typically lacked statistical power. Additionally, it was often difficult to determine whether the subgroup analyses were prespecified or posthoc.

Pharmacogenetics

A rapidly emerging field in medicine is that of pharmacogenetics, which holds promise for better identification of patient groups who may benefit more from a treatment or who are more likely to develop serious adverse events [14]. Until quite recently the focus was on a limited number of candidate genes due to the high cost of genotyping, but as technologies have improved attention has shifted to genome-wide association (GWA) studies of hundreds of thousands or millions of single-nucleotide polymorphisms (SNPs) [15]. This approach and cost-effective whole-genome sequencing technologies allow examination of the whole genome unconstrained by prior hypotheses on genomic structure or function influencing a given trait [16]. Collection of biologic samples at baseline in large, long-term trials has emerged as a rich source for such pharmacogenetic studies. In participants with or without specific genotypes, one would in subgroup analysis compare treatment responses such as serious adverse events.

Genetic determinants of beneficial responses to a treatment are increasingly investigated, especially in cancer. Three cancer drugs, imatinib mesylate, trastuzumab, and gefitinib have documented efficacy in subsets of patients with specific genetic variants, while two others, irinotecan and 6-mercaptopurine, can be toxic in standard doses in other genetically defined subsets of patients [15]. These developments allow treatment to be cost-effective and more efficacious by limiting recommended

use to those likely to benefit. The strength by which common variants can influence the risk determination ranges from a several-fold increased risk compared to those without the variant to a 1,000-fold increase [17].

The identification of new genetic variants associated with serious adverse events is also a critical area of investigation. The goal is to identify through genetic testing those high-risk patients prior to initiation of treatment. A GWA study identified a SNP within the *SLC01B1* gene on chromosome 12 linked to dose-dependent, statin-induced myopathy [18]. Over 60% of all diagnosed myopathy cases could be linked to the C allele of the SNP rs4149056, which is present in 15% of the population. Identification of C allele carriers prior to initiating therapy could reduce myopathy while retaining treatment benefits by targeting this group for lower doses or more frequent monitoring of muscle-related enzymes.

The large sample size requirements, the analytic problem of multiplicity (a genome-wide panel may have over 2.5 million SNPs after imputation), and the need for replications are discussed in Chap. 17.

Changes of Baseline Measurement

Making use of baseline data will usually add sensitivity to a study. For example, an investigator may want to evaluate a new hypoglycemic agent. She can either compare the mean change in Hb_{A1C} from baseline to some subsequent time in the intervention group against the mean change in the control group, or simply compare the mean Hb_{A1C} of the two groups at the end of the study. The former method usually is a more powerful statistical technique because it can reduce the variability of the response variables. As a consequence, it may permit either fewer participants to be studied or a smaller difference between groups to be detected.

Evaluation of possible unwanted reactions requires knowledge – or at least tentative ideas – about what effects might occur. The investigator should record at baseline those clinical or laboratory features which are likely to be adversely affected by the intervention. Unexpected adverse reactions might be missed, but the hope is that animal studies or earlier clinical work will have identified the important factors to be measured.

Natural History Analyses

Baseline measurements enable investigators to perform natural history analyses in a control group which is on either placebo or no uniformly administered intervention. The prognostic importance of suspected risk factors for a variety of fatal and nonfatal events can be evaluated, particularly in large, long-term trials. This evaluation can include verification of previously ascertained risk factors as well as identification of others not earlier considered. Such analyses, although peripheral to the main objectives of a clinical trial, may be important for future research efforts. Their

potential importance is especially true if variables which are subject to intervention can be identified [19]. Even if they are not variables which can be studied in future trials, they can be used in future stratification or subgroup analyses. It should be recognized that trial participants are a selected group defined by the inclusion and exclusion criteria and their willingness to volunteer. Therefore, any natural history analysis from a clinical trial may not be fully generalizable, and is unlikely to substitute for well-designed observational studies.

What Constitutes a True Baseline Measurement?

Screening for Participants

In order to describe accurately the study participants, baseline data should ideally reflect the true condition of the participants. Certain information can be obtained accurately by means of one measurement or evaluation at a baseline interview and examination. However, for many variables, accurately determining the participant's true state is difficult, since the mere fact of impending enrollment in a trial, random fluctuation or the baseline examination itself may alter a measurement. For example, is true blood pressure reflected by a single measurement taken at baseline? If more than one measurement is made, which one should be used as the baseline value? Is the average of repeated measurements recorded over some extended period of time more appropriate? Does the participant need to be taken off all medications or be free of other factors which might affect the determination of a true baseline level? When resolving these questions, the screening required to identify eligible potential participants, the time and cost entailed in this identification, and the specific uses for the baseline information must be taken into account.

In almost every clinical trial, some sort of screening of potential participants is necessary. This may take place over more than one visit. Screening eliminates participants who, based on the entrance criteria, are ineligible for the study. A prerequisite for inclusion is the participant's willingness to comply with a possibly long and arduous study protocol. The participant's commitment, coupled with the need for additional measurements of eligibility criteria, means that intervention allocation usually occurs later than the time of the investigator's first contact with the participant. An added problem may result from the fact that discussing a study with someone or inviting him to participate in a clinical trial may alter his state of health. For instance, people asked to join a study of lipid-lowering agents because they had an elevated serum LDL cholesterol at a screening examination might change their diet on their own initiative just because of the fact they were invited to join the study. Therefore, their serum LDL cholesterol as determined at baseline, perhaps a month after the initial screen, may be somewhat lower than usual. Improvement could happen in many potential candidates for the trial and could affect the validity of the assumptions used to calculate sample size. If the study calls for a special dietary regimen, this might not be so effective at the new, lowered LDL cholesterol level.

As a result of the modification in participant behavior, there may be less room for response to the intervention. Obviously, these changes occur not just in the group randomized to the active intervention, but also in the control group.

Although it may be impossible to avoid altering the behavior of potential participants, in the study design it is often possible to adjust for such anticipated changes. Special care can be taken when discussing studies with people to avoid sensitizing them. Time between invitation to join a study and baseline evaluation can be kept to a minimum. People who have greatly changed their eating habits between the initial screen and baseline, as determined by a questionnaire at baseline, can be declared ineligible to join. Alternatively, they can be enrolled and the required sample size increased. Whatever is done, these are expensive ways to compensate for the reduced expected response to the intervention.

Regression Toward the Mean

Sometimes a person's eligibility for a study is determined by measuring continuous variables, such as blood pressure or cholesterol level. If the entrance criterion is a high or low value, a phenomenon referred to as "regression toward the mean" is encountered [20]. Regression toward the mean occurs because measurable characteristics of an individual do not have constant values but vary. Thus, individuals have days when the measurements are on the high side and other days when they are on the low side within their ranges of variability. Because of this variability, although the population mean for a characteristic may be relatively constant over time, the locations of individuals within the population change. If two sets of measurements are made on individuals within the population, the correlation between the first and second series of measurements will not be perfect. That is, depending on the variability, the correlation will be something less than 1. In addition, it is often the case that the more distant a measured characteristic is from the population mean of that characteristic, the more variable the measurement tends to be.

Therefore, whenever participants are selected from a population on the basis of the cutoff of some measured characteristic, the mean of a subsequent measurement will be closer to the population mean than is the first measurement mean. Furthermore, the more extreme the initial selection criterion (that is, the further from the population mean), the greater will be the regression toward the mean at the time of the next measurement. The "floor-and-ceiling effect" used as an illustration by Schor [21] is helpful in understanding this concept. If all the flies in a closed room are near the ceiling in the morning, than at any subsequent time during the day more flies will be below where they started than above. Similarly, if the flies start close to the floor, the more probable it is for them to be higher, rather than lower, at any subsequent time.

Cutter [22] gives some nonbiological examples of regression toward the mean. He presents the case of a series of three successive tosses of two dice. The average of the first two tosses is compared with the average of the second and third tosses.

If no selection or cut-off criterion is used, the average of the first two tosses would, in the long run, be close to the average of the second and third tosses. However, if a cut-off point is selected, which restricts the third toss to only those instances where the average of the first and second tosses is nine or greater, regression toward the mean will occur. The average of the second and third tosses for this selected group will be less than the average of the first two tosses for this group.

As with the example of the participant changing his diet between screening and baseline, this phenomenon of regression toward the mean can complicate the assessment of intervention. In another case, an investigator may wish to evaluate the effects of an antihypertensive agent. She measures blood pressure once at the baseline examination and enters into the study only those people with systolic pressures over 140 mmHg. She then gives a drug and finds on rechecking that most people have responded with lowered blood pressures. However, when she re-examines the control group, she finds that most of those people also have lower pressures. Regression to the mean is the major explanation for the frequently seen marked mean blood pressure reduction observed early in the control group. The importance of a control group is obvious in such situations. An investigator cannot simply compare preintervention and postintervention values in the intervention group. She must compare postintervention values in the intervention group with values obtained at similar times in the control group.

This regression toward the mean phenomenon can also lead to a problem discussed previously. Because of regression, the values at baseline are less extreme than the investigator had planned on, and there is less room for improvement from the intervention. In the blood pressure example, after randomization, many of the participants may have systolic blood pressures in the low 130s or even below 130 rather than above 140 mmHg. There may be a reluctance to use antihypertensive agents in people with such pressures, and certainly, the opportunity to demonstrate full effectiveness of the agent may be lost.

Two approaches to reducing the impact of regression toward the mean have been used by trials relying on measurements with large variability, such as blood pressure and some chemical determinations. One approach is to use a more extreme value than the entrance criterion when people are initially screened. Secondly, mean values of multiple measurements at the same visit or from more than one screening visit have been used to achieve more stable measurements. In hypertensive trials with a cutoff of systolic blood pressure of 140 mmHg, only those whose second and third measure averaged 150 mmHg or greater would be invited at the first screening visit to the clinic for further evaluation. The average of two recordings at the second visit would constitute the baseline value for comparison with subsequent determinations.

Interim Events

When baseline data are measured too far in advance of intervention assignment, a study event may occur in the interim. The participants having events in the interval

between allocation and the actual initiation of intervention would dilute the results and decrease the chances of finding a significant difference. In the European Coronary Surgery Study, coronary artery bypass surgery should have taken place within 3 months of intervention allocation [23]. However, the mean time until surgery was 3.9 months. Consequently, of the 21 deaths in the surgical group in the first 2 years, six occurred before surgery could be performed. If the response, such as death, is nonrecurring and this occurs between baseline and the start of intervention, the number of participants at risk of having the event later is reduced. Therefore, the investigator needs to be alert to any event occurring after baseline but before intervention is instituted. When such an event occurs before randomization, i.e., allocation to intervention or control, she can exclude the participant from the study. When the event occurs after allocation, but before start of intervention, participants should nevertheless be kept in the study and the event counted in the analysis. Removal of such participants from the study may bias the outcome. For this reason, the European Coronary Surgery Study Group kept such participants in the trial for purposes of analysis. The appropriateness of withdrawing participants from data analysis is discussed more fully in Chap. 17.

Uncertainty About Qualifying Diagnosis

A growing problem in many disease areas such as arthritis, diabetes, and hypertension is finding potential participants who are not receiving treatment for their condition. So-called washout phases are often relied on in order to determine “true” baseline values.

Particularly difficult are those studies where baseline factors cannot be completely ascertained until after intervention has begun. For optimal benefit of thrombolytic therapy in patients with a suspected acute myocardial infarction, treatment has to be given within hours. This means that there is no time to wait for confirmation of the diagnosis with development of Q-wave abnormalities on the ECG and marked increases in serum levels of cardiac enzymes. In the Global Utilization of Streptokinase and Tissue Plasminogen Activator of Occluded Coronary Arteries (GUSTO), trial treatment had to be given within 6 h [24]. To confirm the diagnosis, the investigators had to settle for two less definitive criteria; chest pain lasting at least 20 min and ST-segment elevations on the ECG.

The challenge in the National Institute of Neurological Disorders and Stroke t-PA stroke trial was to obtain a brain imaging study and to initiate treatment within 180 min of stroke onset. This time was difficult to meet and participant enrollment lagged. As a result of a comprehensive process improvement program at the participating hospitals, the time between hospital admission and treatment was substantially reduced with increased recruitment yield. Almost half of eligible patients admitted within 125 min of stroke onset were enrolled [25].

Even if an investigator can get baseline information just before initiating intervention, she may need to compromise. For instance, being an important prognostic

factor, serum cholesterol level is obtained in most studies of heart disease. Serum cholesterol levels, however, are temporarily lowered during the acute phase of a myocardial infarction. Therefore, in any trial using people who have just had a myocardial infarction, baseline serum cholesterol data relate poorly to their usual levels. Only if the investigator has data on participants from a time before the myocardial infarction would usual cholesterol levels be known. Cholesterol levels obtained at the time of the infarction might not allow her to evaluate the natural history or make reasonable observations about changes in cholesterol that occurred because of the intervention. On the other hand, because she has no reason to expect that one group would have greater lowering of cholesterol at baseline than the other group, such levels can certainly tell her whether the study groups are initially comparable.

Contamination of the Intervention

For many trials of chronic conditions, it can be difficult to find and enroll newly diagnosed patients. To meet enrollment goals, investigators often take advantage of available pools of treated patients. In order to qualify such patients, they often have to be withdrawn from their treatment. The advantage of treatment withdrawal is that a true baseline can be obtained. However, there are ethical issues involved with withdrawing active treatments (Chap. 2).

An alternative may be to lower the eligibility criteria for this group of treated patients. In the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT), treated hypertensive patients were enrolled even if their initial blood pressures were below the study/goal blood pressures [26]. It was assumed that these individuals were truly hypertensive and, thus, had elevated blood pressures prior to being given antihypertensive medications. The disadvantage of this approach is that the true untreated values for blood pressure were unknown at baseline.

Medications that participants are taking may also complicate the interpretation of the baseline data and restrict the uses to which an investigator can put baseline data. Determining the proportion of diabetic participants in a clinical trial based on the number with elevated fasting blood sugar or Hb_{A1C} levels at a baseline examination will underestimate the true prevalence. People treated with oral hypoglycemic agents or insulin may have their laboratory values controlled. Thus, the true prevalence of diabetics would be untreated participants with elevated blood sugar or Hb_{A1C} and those being treated for their diabetes regardless of their laboratory values. Similarly, a more accurate estimate of the prevalence of hypertension would be based on the number of untreated hypertensive subjects at baseline plus those receiving antihypertensive treatment.

Withdrawing treatment prior to enrollment could introduce other potential problems. Study participants with a supply of the discontinued medications left in their medicine cabinet may use them during the trial and thus, contaminate the findings.

Similarly, if they have used other medications prescribed for the condition under study, they may also resort to these, whether or not their use is allowed according to the study protocol. The result may be discordant use in the study groups. Assessing and adjusting for the concomitant drug use during a trial can be complex. The use and frequency of use need to be considered. All of these potential problems are much smaller in trials of newly diagnosed patients.

Appreciating that, for many measurements, baseline data may not reflect the participant's true condition at the time of baseline, investigators perform the examination as close to the time of intervention allocation as possible. Baseline assessment may, in fact, occur shortly after allocation but prior to the actual start of intervention. The advantage of such timing is that the investigator does not spend extra time and money performing baseline tests on participants who may turn out to be ineligible. The baseline examination then occurs immediately after randomization and is performed not to exclude participants, but solely as a baseline reference point. Since allocation has already occurred, all participants remain in the trial regardless of the findings at baseline. This reversal of the usual order is not recommended in single-blind or unblinded studies because it raises the possibility of bias during the examination. If the investigator knows to which group the participant belongs, she may subconsciously measure characteristics differently, depending on the group assignment. Furthermore, the order reversal may unnecessarily prolong the interval between intervention allocation and its actual start.

Assessment of Baseline Comparability

Assessment of baseline comparability is important in all trials, and particularly so in nonrandomized studies. The investigator needs to look at baseline variables in several ways. The simplest is to compare each variable to make sure that it has reasonably similar distribution in each study group. Means, medians, and ranges are all convenient measures. The investigator can also combine the variables, giving each one an appropriate weight or coefficient, but doing this presupposes knowledge of the relative prognostic importance of the variables. This kind of knowledge can come only from another study with a very similar population or by looking at the control group after the present study is completed. The weighting technique has the advantage that it can take into account numerous small differences between groups. If imbalances between most of the variables are in the same direction, the overall imbalance can turn out to be large, even though differences in individual variables are small.

In the 30-center Aspirin Myocardial Infarction Study which involved over 4,500 subjects, each center can be thought of as a small study with about 150 subjects [6]. When the baseline comparability within each center was reviewed, substantial differences in almost half the centers were found, some favoring intervention and some control (Furberg, CD, unpublished data). The difference between intervention and control groups in predicted 3-year mortality, using the Coronary Drug Project

model exceeded 20% in five of the 30 clinics. Therefore, all factors which are known or suspected to be important in the subsequent course of the condition under study should be looked at when interpreting results.

Identified imbalances do not invalidate a randomized trial, but they may make interpretation of results more complicated. In the North American Silver-Coated Endotracheal Tube trial, a higher number of patients with chronic obstructive pulmonary disease were randomized to the uncoated tube group [27]. The accompanying editorial [28] points to this imbalance as one factor behind the lack of robustness of the results, which indicated a reduction in the incidence of ventilator-associated pneumonia. Chronic obstructive pulmonary disease is a recognized risk factor for ventilator-associated pneumonia.

Item 15 of the CONSORT statement recommends that the investigators report "baseline demographics and clinical characteristics of each group" (Chap. 19). This is typically done in the first table of a results paper. The statement does not comment on the need for statistical testing of baseline balances (see below).

Testing for Baseline Imbalance

There is a debate over whether one should do formal statistical testing of baseline imbalances [29–32]. Several authors have gone so far to suggest that such testing be avoided [29, 32]. When comparing baseline factors, remember that groups can never be shown to be identical. Only absence of "significant" differences can be demonstrated. A review of 80 trials published in four leading journals, showed that hypothesis tests of baseline comparability were conducted in 46 of these trials. Of a total of 600 such tests, only 24 (4%) were significant at the 5% level [2], consistent with what would be expected by chance.

Testing baseline factors for imbalance is recommended. Specifically, it is important to know which baseline factors may influence the trial outcomes and to determine whether they were imbalanced and whether observed trends of imbalance favored one group or the other. The critical baseline factors to consider ought to be prespecified in the protocol. Reliance on significance testing as a measure of baseline equivalence is common [1]. Due to the often large number of statistical tests, the challenge is to understand the meaning and importance of observed differences. A nonsignificant baseline group difference in the history of hemorrhagic stroke could still affect the treatment outcome in thrombolytic trials [32].

We recommend that the investigators present in the Results section relevant baseline data with *p*-values, or an asterisk indicating which comparisons are significant at a nominal *p*-value (e.g., 0.05 or 0.01) or *z*-scores (standardized differences) from which a *p*-value can be calculated. They should highlight any clinically important differences that may influence the reported findings even if they don't reach nominal statistical significance. In the Discussion, they ought to comment on the effect of such baseline imbalances on the internal validity, as well as how the study population compares with patients seen in clinical practice.

References

1. Hall JC, Hall JL. Baseline comparisons in surgical trials. *ANZ J Surg* 2002;72:567–569.
2. Altman DG, Doré CJ. Randomization and baseline comparisons in clinical trials. *Lancet* 1990;335:149–153.
3. Lachin JM. Properties of simple randomization in clinical trials. *Control Clin Trials* 1988;9:312–326.
4. Harmatz P, Giugliani R, Schwartz I, et al. Enzyme replacement therapy for mucopolysaccharidosis VI: A phase 3, randomized, double-blind, placebo-controlled, multinational study of recombinant human N-acetylgalactosamine 4-sulfatase (recombinant human arylsulfatase B or RHASB) and follow-on, open-label extension study. *J Pediatr* 2006;148:533–539.
5. Sano M, Ernesto C, Thomas RG, et al. for the Members of the Alzheimer's Disease Cooperative Study. A controlled trial of selegiline, alpha-tocopherol, or both as treatment for Alzheimer's Disease. *N Engl J Med* 1997;336:1216–1222.
6. Aspirin Myocardial Infarction Study Research Group. A randomized, controlled trial of aspirin in persons recovered from myocardial infarction. *JAMA* 1980;243:661–669.
7. The Canadian Cooperative Study Group. A randomized trial of aspirin and sulfinpyrazone in threatened stroke. *N Engl J Med* 1978;299:53–59.
8. Antiplatelet Trialists' Collaboration. Collaborative overview of randomised trials of antiplatelet therapy – I. Prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients. *Br Med J* 1994;308:81–106.
9. Ridker PM, Cook NR, Lee I-M, et al. A randomized trial of low-dose aspirin in the primary prevention of cardiovascular disease in women. *N Engl J Med* 2005;352:1293–1304.
10. Kahn SE, Haffner SM, Heise MA, et al. for the ADOPT Study Group. Glycemic durability of rosiglitazone, metformin, or glyburide monotherapy. *N Engl J Med* 2006;355:2427–2443.
11. Dormandy JA, Charbonnel B, Eckland DJA, et al. Secondary prevention of macrovascular events in patients with type 2 diabetes in the PROactive study (PROspective pioglitAzone Clinical Trial In macroVascular Events): A randomised controlled trial. *Lancet* 2005;366:1279–1289.
12. Loke YK, Singh S, Furberg CD. Long-term use of thiazolidinediones and fractures in type 2 diabetes: a meta-analysis. *CMAJ* 2009;180:32–39.
13. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000;355:1064–1069.
14. Johnson JA, Boerwinkle E, Zineh I, et al. Pharmacogenomics of antihypertensive drugs: Rationale and design of the Pharmacogenomic Evaluation of Antihypertensive Responses (PEAR) study. *Am Heart J* 2009;157:442–449.
15. Grant SF, Hakonarson H. Recent development in pharmacogenomics: from candidate genes to genome-wide association studies. *Expert Rev Mol Diagn* 2007;7:371–393.
16. Donnelly P. Progress and challenges in genome-wide association studies in humans. *Nature* 2008;456:728–731.
17. Nelson MR, Bacanu S-A, Mosteller M, et al. Genome-wide approaches to identify pharmacogenetic contributions to adverse drug reactions. *Pharmacogenomics J* 2009;9:23–33.
18. The SEARCH Collaborative Group. *SLCO1B1* variants and statin-induced myopathy – A genomewide study. *N Engl J Med* 2008;359:789–799.
19. Coronary Drug Project Research Group. Factors influencing long-term prognosis after recovery from myocardial infarction – three year findings of the Coronary Drug Project. *J Chronic Dis* 1974;27:267–285.
20. James KE. Regression toward the mean in uncontrolled clinical studies. *Biometrics* 1973;29:121–130.
21. Schor SS. The floor-and-ceiling effect. *JAMA* 1969;207:120.
22. Cutter GR. Some examples for teaching regression toward the mean from a sampling viewpoint. *Am Stat* 1976;30:194–197.
23. European Coronary Surgery Study Group. Coronary-artery bypass surgery in stable angina pectoris: survival at two years. *Lancet* 1979;1:889–893.

24. The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Engl J Med* 1993;329:673–682. (Correction 1994;331:277).
25. Tilley BC, Lyden PD, Brott TG, et al. for the National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Total quality improvement method for reduction of delays between emergency department admission and treatment of acute ischemic stroke. *Arch Neurol* 1997;54:1466–1474.
26. Davis BR, Cutler JA, Gordon DJ, et al. for the ALLHAT Research Group. Rationale and design of the Antihypertensive and Lipid Lowering treatment to prevent Heart Attack Trial (ALLHAT). *Am J Hypertens* 1996;9:342–360.
27. Kollef MH, Afessa B, Anzueto A, et al. for the NASCENT Investigation Group. Silver-coated endotracheal tubes and incidence of ventilator-associated pneumonia. The NASCENT randomized trial. *JAMA* 2008;300:805–813.
28. Chastre J. Preventing ventilator-associated pneumonia. Could silver-coated endotracheal tubes be the answer? (Editorial). *JAMA* 2008;300:842–844.
29. Senn S. Testing for baseline balance in clinical trials. *Stat Med* 1994;13:1715–1726.
30. Steyerberg EW, Bossuyt PMM, Lee KL. Clinical trials in acute myocardial infarction: Should we adjust for baseline characteristics? *Am Heart J* 2000;139:745–751.
31. Roberts C, Torgerson DJ. Understanding controlled trials. Baseline imbalance in randomised controlled trials. *Br Med J* 1999;319:185.
32. Burgess DC, Gebski VJ, Keech AC. Baseline data in clinical trials. *MJA* 2003;179:105–107.

Chapter 10
Recruitment of Study Participants

Often the most difficult task in a clinical trial involves obtaining sufficient study participants within a reasonable time. Time is a critical factor for both scientific and logistical reasons. From a scientific viewpoint, there is an optimal window of time within which a clinical trial can and should be completed. Changes in medical practice, including introduction of new treatment options, may make the trial outdated before it is completed. Other investigators may answer the questions sooner. In terms of logistics, the longer recruitment extends beyond the initially allotted recruitment periods, the greater the pressure becomes to meet the goal. Lagging recruitment will also reduce the statistical power of the trial. Costs increase, frustration, and discouragement often follow. The primary reasons for recruitment failure include overoptimistic expectations, failure to start on time, inadequate planning, and insufficient effort.

Approaches to recruitment of participants will vary depending on the type and size of the trial, the length of time available, the setting (hospital, physician's office, community), whether the trial is single- or multicenter, and many other factors. Because of the broad spectrum of possibilities, this chapter summarizes concepts and general methods rather than elaborating on specific techniques. Emphasis is placed on anticipating and preventing problems. This chapter addresses plans for the recruitment effort, common recruitment problems, major recruitment strategies and sources, actual conduct, and monitoring of recruitment.

Fundamental Point

Successful recruitment depends on developing a careful plan with multiple strategies, maintaining flexibility, establishing interim goals, preparing to devote the necessary effort and obtaining the sample size in a timely fashion.

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_10, © Springer Science+Business Media, LLC 2010

Considerations Before Participant Enrollment

Selection of Study Sample

In Chap. 4, we define the study population as "the subset of the population with the condition or characteristics of interest defined by the eligibility criteria." The group of participants actually recruited into the trial, i.e., the study sample, is a selection from this study population. Those enrolled into a trial do not represent a random sample of those eligible for enrollment. Eligible individuals who volunteer to participate in a randomized trial may be different from eligible nonparticipants (see below). The impact of this potential selection bias on the results of a trial is not well understood. A better understanding of the factors that influence either willingness or unwillingness to participate in a research project can be very helpful in the planning of recruitment efforts.

A thorough literature review through 2001 identified 14 studies that had addressed the question – What reasons do people give for participating and not participating in clinical trials? [1] The answers came from 2,189 participants and 6,498 who declined. The variability was large, but trial participants gave as their major reason for participating potential health benefit (45%), physician influence (27%), and potential benefit to others (18%). Less commonly mentioned reasons given by participants in other studies included a desire to learn more about their condition, get free and better care, encouragement by family members and friends, favorable impression of and trust in clinical staff and even to help promote the investigators' careers [2–5].

Several reasons for declining participation in research projects have also been reported. In the ECRI survey, the major general reasons for not participating were inconvenience (25%), concern over experimentation (20%), potential lack of health benefit (19%), and physician influence (14%). Many patients also lacked interest and preferred to stay with their own physicians. In another survey, fear was given as a major reason by half of those declining participation and the use of a placebo by almost one quarter [5].

Logistical issues are sometimes given – demands on time, conflicts with other commitments, problems with travel/transportation and parking. Barriers to participation in cancer trials include concerns with the trial setting, a dislike of randomization, presence of a placebo or no-treatment group, and potential adverse events [6].

Common Recruitment Problems

The published experience from recruitment of participants into clinical trials through 1995 is nicely summarized in a literature review and annotated bibliography [7]. Over 4,000 titles were identified and 91 articles considered useful for formulation of recruitment strategies in clinical trials are annotated. The literature review

focuses on experiences recruiting diverse populations such as ethnic minorities, women, and the elderly. Also discussed are successful recruitment approaches, which include use of registries, occupational sites, direct mailing, and use of media. The article highlights the value of pilot studies, projecting and monitoring recruitment, and the use of data tracking systems. Many of these issues are covered in more detail later in this chapter.

A review from the UK of 114 clinical trials that recruited participants between 1994 and 2002 explored the factors related to good and poor recruitment [8]. Approximately one-third of all trials met their original recruitment goal within the proposed time frame while approximately half had to be extended. Among those failing to make the original target, one half revised the goals. About 40% of all trials did not initiate recruitment as planned, mostly due to staffing and logistical issues. Almost two-thirds of the trials acknowledged early recruitment problems. More than half of the reviewed trials, a remarkably high number, had a formal pilot study that led to changes in the recruitment approach for the main trial. The written trial materials were revised, the trial design altered, the recruitment target changed, the number of sites increased, and/or the inclusion criteria broadened. A systematic review of recruitment methods identified 14 trials describing 20 different interventions [9]. Strategies that increased recruitment rates were: using an "open" rather than placebo control design, making trial material culturally sensitive, and using telephone reminders and monetary incentives. The authors called for more trials testing interventions to improve trial recruitment.

Even when carefully planned and perfectly executed, recruitment may still proceed slowly. Investigators should always expect problems to occur despite their best efforts. Most of the problems are predictable but a few may be completely unforeseen. In one multicenter study, there were reports of murders of inpatients at the hospital adjacent to the study clinic. It is hardly surprising that attendance at the clinic fell sharply.

Overestimation of eligible participants is a common reason for recruitment difficulties. A group of Finnish investigators [10] conducted a retrospective chart review. The typical eligibility criteria for clinical trials of patients with gastric ulcer were applied to 400 patients hospitalized with that diagnosis. Only 29% met the eligibility criteria but almost all deaths and serious complications such as gastric bleeding, perforation and stenosis during the first 5–7 years occurred among those who would have been ineligible. Clearly, the testing of H_2 blockers or other compounds for the prevention of long-term complication of gastric ulcer in low-risk participants should not be generalized to the entire ulcer population. Troubling in this report is the evidence that the eligibility criteria can have such a dramatic effect of the event rates in those qualifying for participation.

Reliance on physician referrals is common and often problematic. Usually this technique results in very few eligible participants. A survey of 7,000 physicians in 2005 reported that only 31% of them had ever referred a patient to a clinical trial [5]. In one multicenter trial, an investigator invited internists and cardiologists from a large metropolitan area to a meeting. He described the study, its importance and his need to recruit men who had had a myocardial infarction. Each of the physicians

stood up and promised to contribute one or more participants. One hundred fifty participants were pledged; only five were ultimately referred. Despite this, such pleas may be worthwhile because they make the professional community aware of a study and its purpose. Investigators who stay in close contact with physicians in a community and form a referral network have more success in obtaining cooperation and support.

When recruitment becomes difficult, one possible outcome is that an investigator will begin to interpret loosely entry criteria or will deliberately change data to enroll otherwise ineligible participants or even "enroll" fictitious subjects. Unfortunately, this issue is not merely theoretical. Such practices have occurred, to a limited extent, in more than one trial [11–13]. The best way to avoid the problem is to make it clear that this type of infraction harms both the study and the participants, and that neither science nor the investigators are served well by such practices. An announced program of random record audits by an independent person or group during the trial may serve as a deterrent.

Planning

In the planning stage of a trial, an investigator needs to evaluate the likelihood of obtaining sufficient study participants within the allotted time. This planning effort entails obtaining realistic estimates of the number of available potential participants meeting the study entry criteria. However, in the USA, access to available patient data from paper and electronic medical records requires compliance with the Health Insurance Portability and Accountability Act (HIPAA) and similar regulations apply in many other countries. Access can be granted but many community practices do not have such a mechanism in place and tend to be reluctant to release patient information. Even if those restrictions are overcome, census tract data or hospital and physician records may be out of date, incomplete, or incorrect. People may have moved or died since the records were last updated. Information about current use of drugs or frequency of surgical procedures may not reflect what will occur in the future, when the trial is actually conducted. Records may not give sufficient – or even accurate – details about potential participants to determine the impact of all exclusion criteria. Clearly, available data certainly do not reflect the willingness of people to enroll in the trial or comply with the intervention.

After initial record review, an investigator may find it necessary to expand the population base by increasing the geographical catchment area, canvassing additional hospitals, relaxing one or more of the study entrance criteria, increasing the planned recruitment time, or by combining some of these factors. The preliminary survey of participant sources should be as thorough as possible, since these determinations are better made before, rather than after, a study begins.

Investigator commitment is key to success. A concern is that investigators keep adding new trials to those they already have committed to. Trials with higher payments seem to get more attention. The investigator also needs strong support from

his institution and colleagues. Other investigators in the same institution or at nearby institutions may compete for similar participants. Since participants should generally not be in more than one trial at a time, competing studies may decrease the likelihood that the investigator will meet his recruitment goal. Competition for participants may necessitate reappraising the feasibility of conducting the study at a particular site.

Announcements of the trial should precede initiation of recruitment. The courtesy of informing area health professionals about the trial in advance can facilitate cooperation, reduce opposition, and avoid local physicians' surprise at first hearing about the study from their patients rather than from the investigator. Talks to local professional groups are critical, but these and any notices regarding a trial should indicate whether the investigator is simply notifying physicians about the study or is actively seeking their assistance in recruiting participants.

Planning also involves setting up a clinic structure for recruitment with interested and involved coinvestigators, an experienced and organized coordinator in charge of recruitment and other staff required for and dedicated to the operations. A close working relationship between the clinic staff and the investigators with regular clinic meetings is crucial from the very beginning to enrollment of the last participant. Careful planning and clear delineation of staff responsibilities are essential features of well-performing recruitment units.

Although recruitment is often expected to be curvilinear, the calculation of a sample size estimate typically assumes a constant rate of enrollment. A slow start can reduce the statistical power of the trial by reducing the average participant follow-up time. Thus, recruitment should begin no later than the first day of the designated recruitment period. As important as the best planning is, commitment and willingness by everyone to spend a considerable amount of time in the recruitment effort are equally important. Just as investigators usually overestimate the number of participants available, they often underestimate the time and effort needed to recruit. Investigators must accommodate themselves to the schedules of potential participants, many of whom work. Thus, recruitment is often done on weekends and evenings, as well as during usual working hours.

The need for multiple recruitment strategies has been well documented [14, 15]. The first randomization should take place on the first day of the identified recruitment period. Therefore, if there is a lengthy prerandomization screening period, adjustments in the timing of the first randomization should be made. Because it is difficult to know which strategies will be productive, it is important to monitor effort and yield of the various strategies. A successful strategy in one setting does not guarantee success in another. The value of multiple approaches is illustrated by one large study in which the investigator identified possible participants and wrote letters to them, inviting them to participate. He received a poor response until his study was featured on local radio and television news. The media coverage had apparently "legitimized" the study as well as primed the community for acceptance of the trial.

Contingency plans must be available in case recruitment lags. Experience has shown that recruitment yields, in general, are much lower than anticipated. Thus, the identified

sources needed to be much larger than the recruitment goals. Hence, additional sources of potential study participants should be kept in reserve. Approval from hospital staff, large group practices, managed care organizations, corporation directors or others controlling large numbers of potential participants often takes considerable time. Waiting until recruitment problems appear before initiating such approval can lead to weeks or months of inaction and delay. Therefore, it is advisable to make plans to use other sources before the study actually gets underway. If they are not needed, little is lost except for additional time used in planning. Most of the time, these reserves will prove useful.

If data concerning recruitment of potential participants to a particular type of trial are scanty, a pilot or feasibility study may be worthwhile. Pilot studies can provide valuable information on optimal participant sources, recruitment techniques, and estimates of yield. In a trial of elderly people, the question arose whether those in their 70s or 80s would volunteer and actively participate in a long-term, placebo-controlled trial. Before implementing a costly full-scale trial, a pilot study was conducted to answer these and other questions [16]. The study not only showed that the elderly were willing participants but also provided information on recruitment techniques. The success of the pilot led to a full-scale trial.

Recruitment Sources

The sources for recruitment depend on the features of the study population; sick people vs. well, hospitalized vs. not, or acute vs. chronic illness. For example, enrollment of acutely ill hospitalized patients can only be conducted in an acute care setting, whereas enrollment of healthy asymptomatic individuals with certain characteristics or risk factors requires a community-based screening program. Following the introduction of the HIPAA and other privacy regulations, readily available sources for recruitment have changed. Identification of potential participation through review of hospital charts is no longer an effective alternative, except through the active involvement of those patients' own physicians. Thus, focus has shifted to direct participant appeal.

Direct invitation to study participants is an appealing approach, since it avoids many confidentiality issues. Solicitation may be done through mass media, wide dissemination of leaflets advertising the trial, or participation by the investigator in health fairs or similar vehicles. None of these methods is foolproof. The yield is often unpredictable and seems to depend predominantly on the skill with which the approach is made and the size and kind of audience it reaches. One success story featured a distinguished investigator in a large city who managed to appear on a local television station's early evening news show. Thousands of people volunteered for the screening program following this single 5-min appeal. Experience, however, has shown that most individuals who respond to a media campaign are not eligible for the trial.

The recruitment into the Systolic Hypertension in the Elderly Program (SHEP) was a major undertaking [17]. A total of almost 450,000 screenees were contacted

to enroll 4,736 (1.1%) participants. One of the major recruitment approaches in SHEP was mass mailings. A total of 3.4 million letters were sent by 14 of the SHEP clinics and the overall response rate was 4.3%. Names were obtained from Departments of Motor Vehicles, voter registration lists, health maintenance organizations, health insurance companies, the AARP, and others. Endorsement was obtained from these organizations and groups. Many of them issued the invitations on their own letterheads. Each mailing included a letter of invitation, a standard brochure describing SHEP and a self-addressed stamped return postcard. Experience showed that the response rates varied by mailing list source. It was also clear that clinics with experienced recruitment staff did better than the others.

A US survey of 620 previous trial participants asked where they first learned about the trials [5]. Media, the most common answer, was given by 30%, but 26% said the Internet. Web-based strategies seem to grow in importance although the yield appears to vary by type of trial. Only 14% in the survey first learned of the trial via physician referral.

Participants may also be approached through a third party. Patient organizations such as local chapters of diseases such as autism and multiple sclerosis may be willing to refer members. For example, an investigator may bring the attention of physicians to his study by means of letters, telephone calls, presentations at professional society meetings, notices in professional journals or exhibits at scientific conferences. The hope is that these physicians will identify a potential participant and either notify the investigator or ask the person to call him. As noted earlier, this usually yields few participants. To overcome the problem with physician referral, sponsors are offering financial incentives. The value of this practice has not been properly evaluated but it has raised ethical issues concerning conflict of interest, disclosure to potential participants, and implications for the informed consent process [18].

The recruitment targets have to be adjusted if special subgroups of the population are being recruited. In response to a relative paucity of clinical trial data on women and minorities, the US Congress in 1995 directed the National Institutes of Health to establish guidelines for inclusion of these groups in clinical research. The charge to the Director of NIH to "ensure that the trial is designed and carried out in a manner sufficient to provide valid analysis of whether the variables being studied in the trial affect women and members of minority groups, as the case may be, differently than other subjects in the trial" has major implications depending on the interpretation of the term "valid analysis" [19].

To document a similar effect, beneficial or harmful, separately for both men and women, and separately for various racial/ethic groups could increase the sample size by a factor ranging from 4 to 16. The sample size will grow considerably more if the investigator seeks to detect differences in response among the subgroups. We support adequate representation of women and minorities in clinical trials, but suggest that the primary scientific question being posed be the primary determinant of the composition of the study population and the sample size. When the effort is made, successful enrollment of women and minorities can be accomplished. An example is the Selenium and Vitamin E Cancer Prevention Trial [20].

An increasingly common approach to meeting the need for large sample sizes in multicenter trials with mortality and major event response variables has been to establish clinical centers internationally. This experience has been positive and the number of participants enrolled by such centers often exceeds those in the country of the study's origin. The success in recruitment may, however, come at a cost. The trial findings at international sites, especially those in developing countries may differ from those in the originating country, which is often in the developed world. Possible reasons include differences in the baseline characteristics of the study population, in the practice of medicine as a reflection of the quality of care, research traditions and socioeconomic and other factors [21, 22]. O'Shea and Califf analyzed the international differences in cardiovascular trials and reported important differences in participant characteristics, concurrent therapies, coronary revascularizations, length of hospital stay and clinical outcomes in the U.S. and elsewhere [23]. Importantly, they pointed out that, in general, the differing event rates would not be expected to affect the relative effects of a treatment. This is in contrast to a review of 657 abstracts from trials of acupuncture and other interventions [24]. The authors of that review concluded that some countries published unusually high proportions of positive results. Possible explanations include publication biases, level of care, and differences in study populations.

The issue is – Can findings from developing countries be extrapolated to developed countries and regions and vice versa? It is important that the results papers from large international studies address this question by presenting findings by participating country or broad region and continent.

Conduct

Successful recruitment of participants depends not only on proper planning but also on the successful implementation of the plan. Systems must be in place to identify all potential participants from the identified recruitment pool and to screen these people for eligibility. For hospital-based studies, logging all admissions to special units, wards, or clinics is invaluable. However, keeping such logs complete can be difficult, especially during evenings or weekends. During such hours, those most dedicated to the study are often not available to ensure accuracy and completeness. Vacation times and illness may also present difficulties in keeping the log up to date. Therefore, frequent quality checks should be made. Participant privacy is also important. At what point do the investigators obtain consent? For those who refuse to participate, what happens to the data that had been collected and used to identify them? The answers to this will vary from institution to institution and depend on who is keeping the log and for what reason. Information recorded by code numbers can facilitate privacy. The use of data warehouses can be utilized. Electronic medical records permit software algorithms to search for patient profiles that match a particular protocol and automatically identify for the health care team those eligible for a specific trial.

For community-based studies, screening large numbers of people is typically a major undertaking especially if the yield is low. Prescreening potential participants by telephone to identify those with major exclusion criteria (e.g., using demographics, medical history) has been employed in many projects. In the Lung Health Study, investigators used prescreening to reduce the number of screening visits to approximately half of those projected [25, 26]. Investigators need to identify the best times to reach the maximum number of potential participants. If they intend to make home visits or hope to contact people by telephone, they should count on working evenings or weekends. Unless potential participants are retired, or investigators plan on contacting people at their jobs (which, depending on the nature of the job, may be difficult), normal working hours may not be productive times. Vacation periods and summers are additional slow periods for recruitment.

The logistics of recruitment may become more difficult when follow-up of enrolled participants occurs while investigators are still recruiting. In long-term studies, the most difficult time is usually toward the end of the recruitment phase when the same staff, space, and equipment may be used simultaneously for participants seen for screening, baseline, and follow-up examinations. Resources can be stretched to the limit and beyond if appropriate planning has not occurred.

The actual mechanics of recruiting participants needs to be established in advance. A smooth clinic operation is beneficial to all parties. Investigators must be certain that necessary staff, facilities, and equipment are available at appropriate times in the proper places. Keeping potential participants waiting is a poor way to earn their confidence.

Investigators and staff need to keep abreast of recruitment efforts. Conducting regular staff meetings and generating regular reports may serve as forums for discussion of yields from various strategies, percent of recruitment goal attained as well as brainstorming and morale-boosting. These meetings, useful for both single- and multicenter trials, also provide the opportunity to remind everyone about the importance of following the study protocol including paying careful attention to collection of valid data.

Record keeping of recruitment activities is essential to allow analyses of recruitment yields and costs from the various recruitment strategies. Recruiting large number of potential participants requires the creation of timetables, flowcharts, and databases to ensure that screening and recruitment proceed smoothly. Such charts should include the number of people to be seen at each step in the process at a given time, the number and type of personnel and amount of time required to process each participant at each step, and the amount of equipment needed (with an allowance for “down” time). A planned pilot phase is helpful in making these assessments. One positive aspect of slow early recruitment is that the “bugs” in the start-up process can be worked out and necessary modifications made.

Several additional points regarding the conduct of recruitment are worth emphasizing:

First, the success of a technique is unpredictable. What works in one city at one time may not work at the same place at another time – or in another city. Therefore, the investigator needs to be flexible and to leave room for modifications.

Second, investigators working especially with sick participants must maintain good relationships with participants' personal physicians. Physicians disapproving of the study or of the way it is conducted are more likely to urge their patients not to participate.

Third, investigators must respect the families of potential participants. Most participants like to discuss research participation with their family and friends. Investigators should be prepared to spend time reviewing the study with them. If the study requires long-term cooperation from the participant, we encourage such discussions. Anything that increases family support is likely to lead to better recruitment and protocol adherence.

Fourth, recruiting should not be overly aggressive. While encouragement is necessary, excessive efforts to convince, or "arm twist" people to participate could prove harmful in the long run, in addition to raising ethical concerns. One might argue that excessive salesmanship is unethical. Those reluctant to join may be more likely to abandon the study later or be poor adherers to study interventions after randomization. Effective work on adherence begins during the recruitment phase.

Fifth, the recruitment success is closely associated with the investigator's level of commitment.

Monitoring

Successful trial recruitment often depends on establishing short-term and long-term recruitment goals. The investigator should record these goals and make every effort to achieve them. Since lagging recruitment commonly results from a slow start, timely establishment of initial goals is crucial. The investigator should be ready to randomize participants on the first official day of study opening.

The use of weekly and/or monthly interim goals in a long-term study orients the investigator and staff to the short-term recruitment needs of the study. These goals can serve as indicators for lagging recruitment and may help avoid a grossly uneven recruitment pace. Inasmuch as participant follow-up is usually done at regular intervals, uneven recruitment results in periods of peak and slack during the follow-up phase. This threatens effective use of staff time and equipment. Of course, establishing a goal in itself does not guarantee timely participant recruitment. The goals need to be realistic and the investigator must make the commitment to meet each interim goal.

The reasons for falling behind the recruitment goal(s) should be determined. In a multicenter clinical trial, valuable insight can be obtained by comparing results and experiences from different centers. Those clinical sites with the best recruitment performance can serve as "role models" for other sites, which should be encouraged to incorporate other successful techniques into their recruitment schemes. Multicenter studies require a central office to oversee recruitment, to compare enrollment results, to facilitate communication among sites, and to lend

support and encouragement. Frequent feedback to the centers by means of tables and graphs, which show the actual recruitment compared with originally projected goals, are useful tools. Examples are shown in the following figures and table. Figure 10.1 shows the progress of an investigator who started participant recruitment on schedule and maintained a good pace during the recruitment period. The investigator and clinic staff accurately assessed participant sources and demonstrated a commitment to enrolling participants in a relatively even fashion. Figure 10.2 shows the record of an investigator who started slowly but later

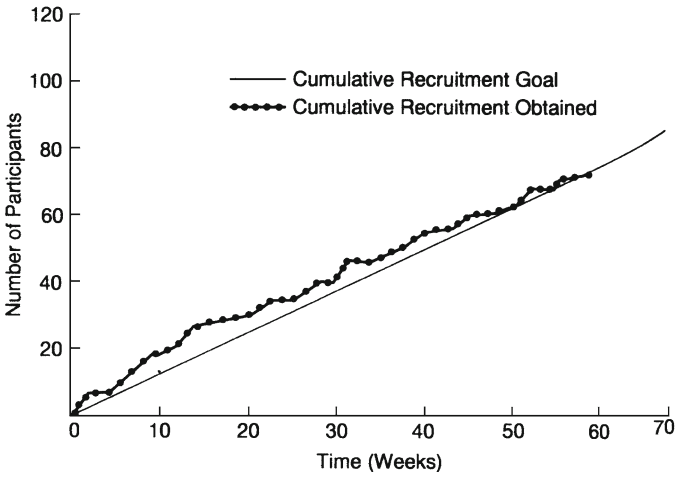


Fig. 10.1 Participant recruitment in a clinic that consistently performed at goal rate

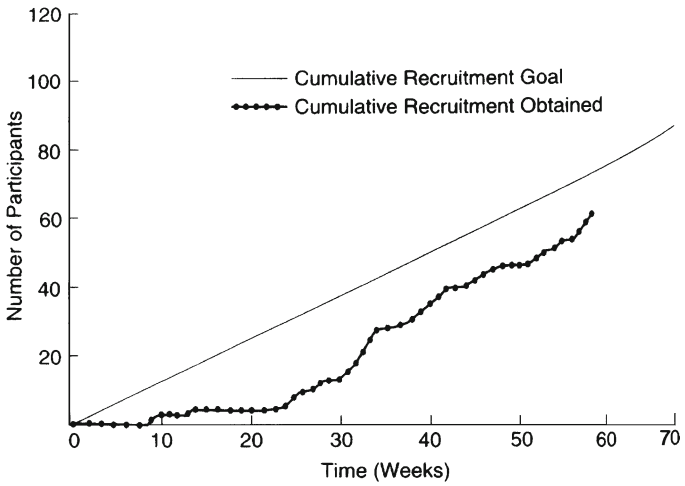


Fig. 10.2 Participant recruitment in a clinic that started slowly and then performed at greater than goal rate

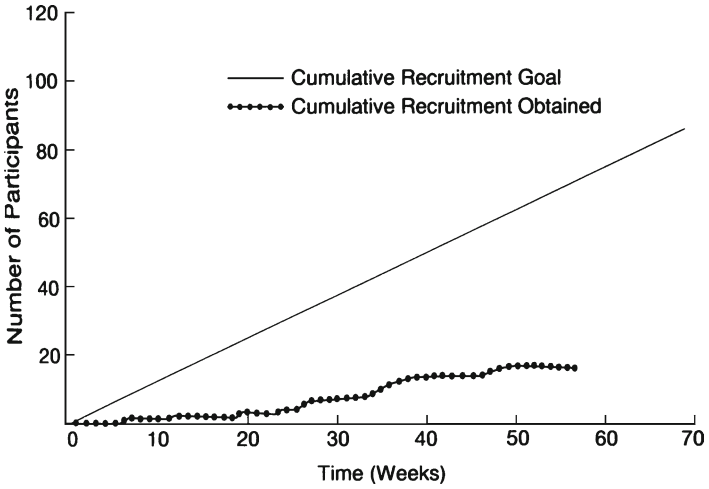


Fig. 10.3 Participant recruitment in a clinic that performed poorly

improved. However, considerable effort was required to compensate for the poor start. Clinic efforts included expanding the base from which participants were recruited and increasing the time spent in enrollment. Even if the clinic eventually catches up, the person-years of exposure to the intervention has been reduced which may affect event rates and trial power. In contrast, as seen in Fig. 10.3, the investigator started slowly and never was able to improve his performance. This center was dropped from a multicenter study because it could not contribute enough participants to the study to make its continued participation efficient.

Table 10.1 shows goals, actual recruitment, and projected final totals (assuming no change in enrollment pattern) for three other centers of a multicenter trial. Such tables are useful to gauge recruitment efforts short term as well as to project final numbers of participants. The tables and figures should be updated as often as necessary.

In single-center trials, the investigator should also monitor recruitment status at regular and frequent intervals. Review of these data with staff keeps everyone aware of recruitment progress. If recruitment lags, the delay can be noted early, the reasons identified and appropriate action taken.

Approaches to Lagging Recruitment

We have identified five possible approaches to deal with lagging recruitment.

The first is to accept a smaller number of participants than originally planned. Doing this is far from ideal, inasmuch as the power of the study will be reduced. In accepting a smaller number of participants than estimated, the investigator must either alter design features such as the primary response variable, or change

Table 10.1 Weekly recruitment status report by center

Center	(1) Contracted goal	(2) Enrollment this week	(3) Actual enrollment to date	(4) Goal enrollment to date	(5) Actual minus goal	(6) Success rate (3)/(4)	(7) Final projected intake	(8) Final deficit or excess (7) − (1)
A	150	1	50	53.4	−3.4	0.94	140	−10
B	135	1	37	48.0	−11.0	0.77	104	−31
C	150	2	56	53.4	2.6	1.06	157	7

Table used in the Beta-Blocker Heart Attack Trial: Coordinating Center, University of Texas, Houston

assumptions about intervention effectiveness and participant adherence. As indicated elsewhere, such changes midway in a trial may be liable to legitimate criticism. Only if the investigator is lucky and discovers that some of the assumptions used in estimating sample size were too pessimistic would this "solution" provide comparable power. There are rare examples of this happening. In a trial of aspirin in people with transient ischemic attacks, aspirin produced a greater effect than initially postulated [27]. Therefore, the less-than-hoped-for number of participants turned out to be adequate. Alternatively, extra effort might be made to achieve better-than-projected participant adherence to the study protocol, and thereby reduce the number of required participants.

A second approach is to relax the inclusion criteria. This should be done only if little expectation exists that the study design will suffer. The design can be marred when, as a result of the new type of participants, the control group event rate is altered to such an extent that the estimated sample size is no longer appropriate. Also, the expected response to intervention in the new participants may not be as great as in the original participants. Furthermore, the intervention might have a different effect or have a greater likelihood of being harmful in the new participants than in those originally recruited. The difference in additional participants would not matter if the proportion of participants randomized to each group stayed the same throughout recruitment. However, as indicated in Chap. 6, certain randomization schemes alter that proportion, depending on baseline criteria or study results. Under these circumstances, changing entrance criteria may create imbalances among study arms.

The Coronary Drug Project provides a classic example [28]. Only people with documented Q-wave myocardial infarctions were originally eligible. With enrollment falling behind, the investigators decided to admit participants with non-Q-wave infarctions. Since there was no reason to expect that the action of lipid-lowering agents that were being studied would be any different in the new group than in the original group and since the lipid-lowering agents were not contraindicated in the new participants, the modification seemed reasonable. However, there was some concern that overall mortality rate would be changed because mortality in people with non-Q-wave infarctions may be less than mortality in people with Q-wave infarctions. Nevertheless, the pressure of recruitment overrode that concern. Possible baseline imbalances did not turn out to be a problem. In this particular study, where

the total number of participants was so large (8,341), there was every expectation that randomization would yield comparable groups. If there had been uncertainty regarding this, stratified randomization could have been employed (Chap. 6). Including people with non-Q-wave infarctions may have reduced the power of the study because this group had a lower mortality rate than those with Q-wave infarctions in each of the treatment groups, including the placebo group. However, the treatments were equally ineffective when people with Q-wave infarctions were analysed separately from people with non-Q-wave infarctions [29].

The third and probably most common approach to recruitment problems is to extend the time for recruitment or, in the case of multicenter studies, to add recruiting centers. Both are the preferred solutions, requiring neither modification of admission criteria nor diminution of power. However, they are also the most costly. Whether the solution of additional time or additional centers is adopted depends on cost, on the logistics of finding and training other high quality centers, and on the need to obtain study results quickly.

A fourth approach to lagging recruitment is "recycling" of potential participants. When a prospective participant just misses meeting the eligibility criteria, the temptation is natural to try to enroll him by repeating a measurement, perhaps under slightly different conditions. Due to variability in a screening test, many investigators argue that it is reasonable to allow one repeat test and give a person interested in the trial a "second chance." In general, this recycling should be discouraged. A study is harmed by enrolling persons for whom the intervention might be ineffective or inappropriate. However, in some progressive diseases, waiting a year to recycle a potential participant may prove to be useful.

Instances exist where, in order to enter a drug study, the participant needs to be off all other medication with similar actions. At baseline, he may be asked whether he has adhered with this requirement. If he has not, the investigator may repeat the instructions and have the participant return in a week for repeat baseline measurements. The entrance criterion checks on a participant's ability to adhere with a protocol and his understanding of instructions. This "second chance" is different from recycling and it is legitimate from a design point of view. However, the second-chance participant, even if he passes the repeat baseline measurement, may not be as good a candidate for the study as someone who adhered on the first occasion [30].

The fifth approach – broadening or changing the prespecified primary response variable is very common. Broadening the prespecified response variable during the conduct of a trial when the observed number of response outcomes, or events, is markedly below what was required ought to be avoided. If done to compensate for low statistical power, the change should be clearly acknowledged in all results papers. The credibility and acceptance of the trial findings are likely to suffer. Changing the primary outcome may have some hazards. The placebo-controlled Carvedilol Post-Infarct Survival Control in Left Ventricular Dysfunction (CAPRICORN) study experienced slow recruitment and a lower than expected event rate. The prespecified primary endpoint was all-cause mortality. The blinded Steering Committee discussed the possibility of adding a second primary endpoint, all-cause mortality or cardiovascular hospitalization. Other solutions were considered but the option of adding a

second primary outcome was recommended. It was decided to allocate a *p*-value of 0.045 to the new endpoint and a *p*-value of 0.005 to the original one in order to maintain an overall type 1 error of 0.05 [31]. At the conclusion of the trial, all-cause mortality had a hazard ratio of 0.77 (95% CI 0.60–0.98) for a $p=0.031$. The hazard ratio for the new endpoint was 0.92 (95% CI 0.80–1.07) for a $p=0.296$. Thus, CAPRICORN would have shown a significant benefit if the primary outcome had not been changed or if the allocation of the *p*-value had emphasized all-cause mortality more. Due to the very favorable outcome of other trials of carvedilol, the results of CAPRICORN were ultimately accepted as positive. However, we strongly discourage changing the primary outcome after a trial is underway.

References

1. ECRI Health Technology Assessment Information Service. Patients' reasons for participation in clinical trials and effect of trial participation on patient outcomes. ECRI Evidence Report. April 2002, Issue 74.
2. Wright JR, Crooks D, Ellis PM, et al. Factors that influence the recruitment of patients to phase III studies in oncology. The perspective of the clinical research assistant. *Cancer* 2002;95:1584–1591.
3. Cox K, McGarry J. Why patients don't take part in cancer clinical trials: an overview of the literature. *Eur J Cancer Care* 2003;12:114–122.
4. Sharp L, Cotton SC, Alexander L, et al. on behalf of the TOMBOLA group. Reasons for participation and non-participation in a randomized controlled trial: postal questionnaire surveys of women eligible for TOMBOLA (Trial of Management of Borderline and Other Low-grade Abnormal smears). *Clin Trials* 2006;3:431–442.
5. Barnes K. Patients provide insight into trial participation. Outsourcing-Pharma.com, July 4, 2007. www.outsourcing-pharma.com/content/view/print/135930.
6. Mills EJ, Seely D, Rachlis B, et al. Barriers to participation in clinical trials of cancer: a meta-analysis and systematic review of patient-reported factors. *Lancet Oncol* 2006;7:141–148.
7. Lovato LC, Hill K, Hertert S, et al. Recruitment for controlled clinical trials: literature summary and annotated bibliography. *Control Clin Trials* 1997;18:328–357.
8. McDonald AM, Knight RC, Campbell MK, et al. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. *Trials* 2006;7:9.
9. Watson JM, Torgerson DJ. Increasing recruitment to randomised trials: a review of randomised controlled trials. *BMC Med Res Methodol* doi:10.1186/1471-2288-6-34 (published 19 July 2006).
10. Kääriäinen I, Sipponen P, Siurala M. What fraction of hospital ulcer patients is eligible for prospective drug trials? *Scand J Gastroenterol* 1991;186:73–76.
11. Sheldon T. Dutch neurologist found guilty of fraud after falsifying 438 case records. *Br Med J* 2002;325:734.
12. Ross DB. The FDA and the case of Ketek. *N Engl J Med* 2007;356:1601–1604.
13. POISE Study Group. Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. *Lancet* 2008;371:1839–1847 (Web attachment 1).
14. Hunninghake DB. Summary conclusions. *Control Clin Trials* 1987;8:1S–5S.
15. Kingry C, Bastien A, Booth G, et al. for the ACCORD Study Group. Recruitment strategies in the Action to Control Cardiovascular Risk in Diabetes (ACCORD) Trial. *Am J Cardiol* 2007;99(Suppl):68i–79i.
16. Hulley SB, Furberg CD, Gurland B, et al. Systolic Hypertension in the Elderly Program (SHEP): antihypertensive efficacy of chlorthalidone. *Am J Cardiol* 1985;56:913–920.

17. Cosgrove N, Borhani NO, Bailey G, et al. Mass mailing and staff experience in a total recruitment program for a clinical trial: the SHEP experience. *Control Clin Trials* 1999;19:133–148.
18. Bryant J, Powell J. Payment to healthcare professionals for patient recruitment to trials: a systematic review. *Br Med J* 2005;331:1377–1378.
19. Freedman LS, Simon R, Foulkes MA, et al. Inclusion of women and minorities in clinical trials and the NIH Revitalization Act of 1993 – the perspective of NIH clinical trialists. *Control Clin Trials* 1995;16:277–285.
20. Cook ED, Moody-Thomas S, Anderson KB, et al. Minority recruitment to the Selenium and Vitamin E Cancer Prevention Trial (SELECT). *Clin Trials* 2005;2:436–442.
21. Shibata M, Flather M, de Arenaza DP, et al. Potential impact of socioeconomic differences on clinical outcomes in international clinical trials. *Am Heart J* 2001;1411:1019–1024.
22. Orlandini A, Diaz R, Wojdyla D, et al. Outcomes of patients in clinical trials with ST-segment elevation myocardial infarction among countries with different gross national incomes. *Eur Heart J* 2006;27:527–533.
23. O'Shea JC, Califf RM. International differences in cardiovascular clinical trials. *Am Heart J* 2001;141:866–874.
24. Vickers A, Goyal N, Harland R, Rees R. Do certain countries produce only positive results? A systematic review of controlled trials. *Control Clin Trials* 1998;19:159–166.
25. Durkin DA, Kjelsberg MO, Buist AS, et al. Recruitment of participants in the Lung Health Study, I: description of methods. *Control Clin Trials* 1993;14:20S–37S.
26. Daly M, Seay J, Balshem A, et al. Feasibility of a telephone survey to recruit health maintenance organization members into a tamoxifen chemoprevention trial. *Cancer Epidemiol Biomarkers Prev* 1992;1:413–416.
27. Fields WS, Lemak NA, Frankowski RF, et al. Controlled trial of aspirin in cerebral ischemia. *Stroke* 1977;8:301–314.
28. The Coronary Drug Project Research Group. The Coronary Drug Project: design, methods, and baseline results. *Circulation* 1973;47:I-1–I-50.
29. The Coronary Drug Project Research Group. Clofibrate and niacin in coronary heart disease. *JAMA* 1975;231:360–381.
30. Sackett DL. A compliance practicum for the busy practitioner. In Haynes RB, Taylor DW, Sackett DL (eds.). *Compliance in Health Care.* Baltimore: Johns Hopkins University Press, 1979.
31. Julian D. The data monitoring experience in the Carvedilol Post-Infarct Survival Control in Left Ventricular Dysfunction Study: hazards of changing primary outcomes. In DeMets DL, Furberg CD, Friedman LM (eds.). *Data Monitoring in Clinical Trials*. New York: Springer, 2006, pp. 337–345.

Chapter 11
Data Collection and Quality Control

No study is better than the quality of its data. Data in clinical trials are collected from several sources – interviews, questionnaires, participant examinations, or laboratory determinations. Also, data that have been collected and evaluated by someone outside the study may be used in a trial; for example, diagnoses obtained from death certificates or hospital records.

Avoiding problems in the data collection represents a challenge. There are many reasons for poor quality data and avoiding them all is difficult, so the goal is to limit their amount and, thus, their impact on the trial findings. Many steps can be taken during the planning phase to optimize collection of high quality data. The problems encompass missing data, erroneous (including falsified and fabricated) data, large variability and long delays in data submission. Even with the best planning, data quality needs to be monitored throughout the trial and corrective actions taken to deal with unacceptable problems. This chapter has sections addressing the problems in data collection, how to minimize collection of poor quality data, and the need for quality monitoring, which includes audits.

Concerted efforts to improve data quality in clinical trials have increased markedly over the past decade or so. The International Conference of Harmonisation (ICH) Good Clinical Practices (GCP) guidelines defined the international ethical and scientific standards for clinical trials in 1996 [1]. They cover the spectrum of phases from design, conduct, recording to reporting. This roadmap of responsibilities has been updated and the latest revision issued in 2007 [2]. Other organizations followed in the ICH's footsteps and have issued their own versions of quality assurance guidelines. The Society for Clinical Trials issued, in 1998, guidelines for multicenter trials [3]. The industry perspective has been reviewed by Williams [4]. The oncology community has guidelines issued by the American Society of Clinical Oncology [5] and special standards for pediatric oncology [6]. There is also a report that resulted from a 2007 Conference on Sensible Guidelines [7]. An article by Acosta et al. [8] discusses the implementation of GCP guidelines in developing countries. Finally, the texts by McFadden [9] and Meinert [10] contain detailed descriptions of data collection.

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_11, © Springer Science+Business Media, LLC 2010

Fundamental Point

During all phases of a study, sufficient effort should be spent to ensure that all data critical to the interpretation of the trial, i.e., those relevant to the main questions posed in the protocol, are of high quality.

The definition of key data depends on trial type and objectives. Baseline characteristics of the enrolled participants, particularly those related to major eligibility measures are clearly key as are primary and secondary outcome measures. The effort expended on assuring freedom from error for key data will be considerable. It is essential that conclusions or interferences from the trial be based on accurate and valid data. Fastidious attention to all data is not possible, nor is it necessary. One approach is to decide in advance the degree of error one is willing to tolerate for each type of data. The key data, as well as certain process information such as informed consent, should be as close to error free as possible. One may be willing to tolerate a greater error rate for other data. The confirmation, duplicate testing, and auditing that is done on data of secondary importance need not be as extensive.

Perhaps only a sampling of audits is necessary.

The data collected should focus on the answers to the questions posed in the protocol. Essential data include the following:

- Baseline information
- Measures of adherence to the study intervention
- Concomitant interventions
- Primary response variable(s)
- Secondary response variables
- Other prespecified response variables
- Adverse events with emphasis on serious events

Data are collected to answer questions about benefit, risk, and ability to adhere to the intervention being tested. Trials must collect data on baseline covariates or risk factors for at least three purposes: (1) to verify eligibility and describe the population studied; (2) to verify that randomization did balance the important known risk factors; and (3) to allow for limited subgroup analyses. Obviously, data must be collected on the primary and secondary response variables specified in the protocol and, in some cases, tertiary level variables. Some measures of adherence to the interventions specified in the protocol are necessary as well as important concomitant medications used during the trial. That is, to test validly of the intervention, the trial must describe how much of the intervention the participant was exposed to, and what other interventions were used. Collection of adverse events is challenging for many reasons (see Chap. 12).

Each data element considered should be examined as to its importance in answering the questions. Trialists cannot include every outcome that might be "nice to know." Each data element requires collection, processing, and quality control, as discussed below, and adds to the cost and the overall burden of the trial.

We think that far too much data are generally collected. Only a small portion are actually used in trial monitoring and publications. Excessive data collection is not only costly but can also indirectly affect the quality of the more critical data elements.

Problems in Data Collection

Major Types

Problems in data collection can be of several sorts and can apply to the initial acquisition of data such as physical examination as well as to the recording of the data on a form or data entry into a remote computer terminal or microcomputer. There are four major types of data problems that are discussed here: (1) missing data, (2) incorrect data, (3) excess variability, and (4) delayed submission.

First, incomplete and irretrievably missing data can arise, for example, from the inability of participants to provide necessary information, from inadequate physical examinations, from laboratory mishaps, from carelessness in completion of study forms or data entry or from inadequate quality control within electronic data management systems. The percent of missing data in a study is considered as one indicator of the quality of the data and therefore, the quality of the trial.

Second, erroneous data may not be recognized and therefore can be even more troublesome than incomplete data. For study purposes, a specified condition may be defined in a particular manner. A clinic staff member may unwittingly use a clinically acceptable definition, but one that is different from the study definition. Specimens may be mislabeled. In one clinical trial, the investigators appropriately suspected mislabeling errors when, in a glucose tolerance test, the fasting glucose levels were higher than the 1 h glucose levels in some participants. Badly calibrated equipment can be a source of error. In addition, the incorrect data may be entered on a form. A blood pressure of 84/142 mmHg, rather than 142/84 mmHg, is easy to identify as wrong. However, while 124/84 mmHg may be incorrect, it is perfectly reasonable, and the error would not necessarily be recognized. The most troublesome types of erroneous data are those that are falsified or entirely fabricated. The pressure to recruit participants may result in alterations of laboratory values, blood pressure measurements, and critical dates in order to qualify otherwise ineligible participants for enrollment [11, 12].

The third problem is variability in the observed characteristics. Variability reduces the opportunity to detect real changes. The variability between repeated assessments can be unsystematic (or random), systematic, or a combination of both. Variability can be intrinsic to the characteristic being measured, the instrument used for the measurement, or the observer responsible for obtaining the data. People can show substantial day-to-day variations in a variety of physiologic measures. Learning effects associated with many performance tests also contribute to variability.

The problem of variability, recognized many decades ago, is not unique to any specific field of investigation [13, 14]. Reports of studies of repeat chemical determinations, determinations of blood pressure, physical examinations, and interpretations of X-rays, electrocardiograms and histological slides, etc. indicate the difficulty in obtaining highly reproducible data. People perform tasks differently and may vary in knowledge and experience. These factors can lead to inter-observer variability. In addition, inconsistent behavior of the same observer between repeated measurements may also be much greater than expected, though intra-observer inconsistency is generally less than inter-observer variability.

Reports from studies of laboratory determinations illustrate that the problem of variability has persisted for at least six decades. In 1947, Belk and Sunderman [15] reviewed the performance of 59 hospital laboratories on several common chemical determinations. Using prepared samples, they found that unsatisfactory results outnumbered the satisfactory. Regular evaluation of method performance, often referred to as proficiency testing, is now routinely conducted and required by laboratories in many countries [16, 17]. All laboratories performing measurements for clinical trials should be Clinical Laboratory Improvement Amendments (CLIA) or similarly approved (http://www.cms.hhs.gov/CLIA/09_CLIA_Regulations_and_Federal_Register_Documents.asp).

Diagnostic procedures that rely on subjective interpretations are not surprisingly more susceptible to variability. One example is radiologists' interpretation of screening mammograms [18]. Nine radiologists read cases with verified cancers, benign, and negative findings in the clinic. Approximately 92% of the mammograms of verified cases were, on an average, read as positive. The reading of the negative mammograms showed a substantial inter-reader variability.

The intra- and interreader variability in QT interval measurement on electrocardiograms was estimated by two different methods [19]. Eight readers analyzed the same set of 100 electrocardiograms twice 4 weeks apart. Five consecutive complexes were measured. For the more commonly used threshold method, the intra-reader standard deviation was 7.5 ms and the inter-reader standard deviation 11.9 ms. Due to the association between QT prolongation and malignant arrhythmias, the U.S. Food and Drug Administration (FDA) is concerned about drugs that prolong the QT interval by a mean of about 5 ms. Thus, the usual variability in measurement is greater than what is considered a clinically important difference.

Another type of variability is the use of nonstandardized terms. As a result, the ability to exchange, share, analyze, and integrate clinical trial data is limited by this lack of coordination in terms of semantics. Increased attention has been devoted to so-called harmonized semantics [20, 21]. A new strategy for international classification and coding of prescription and over-the-counter medications, traditional herbal medicines, and dietary supplements has been proposed [22].

The fourth problem, delayed submission of participant data from the clinical site in multicenter trials, used to be a major issue. However, it has decreased markedly with the onset of web-based data entry (see below).

Minimizing Poor Quality Data

General approaches for minimizing potential problems in data collection are summarized below. Most of these should be considered during the planning phase of the trial. Examples in the cardiovascular field are provided by Luepker et al. [23]. In this section, we discuss design of protocol and manual, development of forms, training and certification, pretesting, techniques to reduce variability, and data entry.

Design of Protocol and Manual

The same question can be interpreted in many ways. Clear definitions of entry and diagnostic criteria and methodology are therefore essential. These should be included in the protocol and written so that all investigators and staff can apply them in a consistent manner throughout the trial. Accessibility of these definitions is also important. Even the same investigator may forget how he previously interpreted a question unless he can readily refer to instructions and definitions. A manual of procedures (MOP) should be prepared in every clinical trial. Although it may contain information about study background, design, and organization, the MOP is not simply an expanded protocol. In addition to listing eligibility criteria and response variable definitions, it should indicate how the criteria and variables are determined. The MOP provides detailed answers to all conceivable "how to" questions. Most important, it needs to describe the participant visits – their scheduling and content – in detail. Instructions for filling out forms, performing tasks such as laboratory determinations, drug ordering, storing and dispensing, and compliance monitoring must be clear and complete. Finally, recruitment techniques, informed consent, participant safety, emergency unblinding, the use of concomitant therapy, and other issues need to be addressed. Updates and clarifications usually occur during the course of a study. These revisions should be made available to every staff person involved in data collection.

Descriptions of laboratory methods or imaging techniques and the ways the results are to be reported also need to be stated in advance. In one study, plasma levels of the drug propranolol were determined by using standardized methods. Only after the study ended was it discovered that two laboratories routinely were measuring free propranolol, and two other laboratories were measuring propranolol hydrochloride. A conversion factor allowed investigators to make simple adjustments and arrive at legitimate comparisons. Such adjustments are not always possible.

Development of Forms

Ideally, the study forms should contain all necessary information [10]. If that is not possible, the forms should outline the key information and refer the investigator to

the appropriate page in the MOP. Well-designed forms will minimize errors and variability. Forms should be as short and as well organized as possible, with a logical sequence to the questions. Forms should be clear, with few "write-in" answers. As little as possible should he left to the imagination of the person completing the form. This means, in general, no essay questions. The questions should elicit the necessary information and little else. Questions that are tacked on because the answers would be "nice to know" are rarely analyzed and may distract attention from pertinent questions. In several studies where death is the primary response variable, investigators may have an interest in learning about the circumstances surrounding the death. In particular, the occurrence of symptoms before death, the time lapse from the occurrence of such symptoms until death, and the activity and location of the participant at the time of death have been considered important and may help in classifying the cause of death. While this may be true, focusing on these details has led to the creation of extraordinarily complex forms, which take considerable time to complete. Moreover, questions arise concerning the accuracy of the information because much of it is obtained from proxy sources who may not have been with the participant when she died. Unless investigators clearly understand how these data will be used, simpler forms are preferable.

A comprehensive review of the multitude of issues in the design of study forms is presented by Cook and DeMets [24]. They describe the categories of data typically collected in randomized clinical trials: participant identification and treatment assignment; screening and baseline information; follow-up visits, tests, and procedures; adherence to study treatment; adverse experiences; concomitant medication and interventions; clinical outcomes and participant treatment, follow-up and survival status. Also discussed are mechanisms for data collection and design and review of case report forms.

Training and Certification

It has long been recognized that training sessions for investigators and staff to promote standardization of procedures are crucial to the success of any large study. Whenever more than one person is filling out forms or examining participants, training sessions help to minimize errors. There may be more than one correct way of doing something in clinical practice, but for study purposes, there is only one way. Similarly, the questions on a form should always be asked in the same way. The answer to, "Have you had any stomach pain in the last 3 months?" may be different from the answer to, "You haven't had any stomach pain in the last 3 months, have you?" Even differences in tone or the emphasis placed on various parts of a question can alter or affect the response. Kahn et al. [25] reviewed the favorable impact of training procedures instituted in the Framingham Eye Study. The 2 days of formal training included duplicate examinations, discussions about differences, and the use of a reference set of fundus photographs. Neaton et al. [26] concluded that initial training is useful and should cover the areas of clinic operation, technical

measurements, and delivery of intervention. Centralized interim training of new staff is less efficient and can be substituted by regional training, teleconferencing, or internet-based approaches.

Mechanisms to verify that all clinic staff do things the same way should be developed. These include instituting certification procedures for specified types of data collection. If blood pressure, electrocardiograms, pulmonary function tests, or laboratory tests are important, the people performing these determinations should not only be trained, but also be tested and certified as competent. Periodic retraining and certification are especially useful in long-term studies since people tend to forget, and personnel turnover is common. For situations where staff must conduct clinical interviews, special training procedures to standardize the approach have been used. In a study of B-mode ultrasonography of the carotid arteries, marked differences in intimal-medial thickness measurements were found between the 13 readers at the reading center [27]. During the 5-year study, the relative biases of readers over time varied, in some cases changing from low to high and vice versa. A sharp increase in average intimal-medial thickness measurements observed toward the end of the study was explained by readers reading relatively high having an increased workload, the hire of a new reader also reading high, and a reader changing from reading low to high.

Pretesting

Pretesting of forms and procedures is almost always essential. Several people similar to the intended participants may participate in simulated interviews and examinations to make sure that procedures are properly performed and questions on the forms flow well and provide the desired information. Furthermore, by pretesting, the investigator and staff grow familiar and comfortable with the form. Fictional case histories can be used to check form design and the care with which forms are completed. When developing forms, most investigators cannot even begin to imagine the numerous ways questions can be misinterpreted until several people have been given the same information and asked to fill out the same form. Part of the reason for different answers is undoubtedly due to carelessness by the person completing the form. The use of “de-briefing” in the pilot test may bring to light misinterpretations that would not be detected when real participants fill out the forms. Inadequacies in form structure and logic can also be uncovered by the use of pretesting. Thus, pretesting reveals areas where forms might be improved and where additional training might be worthwhile.

De-briefing is an essential part of the training process. This helps people completing the forms to understand how the forms are meant to be completed and what interpretations are wanted. Discussion also alerts them to carelessness. When done before the start of the study, this sort of discussion allows the investigator to modify inadequate items on forms. These case history exercises might be profitably repeated several times during the course of a long-term study to indicate when

education and retraining are needed. Ideally, forms should not be changed after the study has started. Inevitably, though, modifications are made. Pretesting can help to minimize them.

Techniques to Reduce Variability

Both variability and bias in the assessment of response variables should be minimized through repeat assessment, blinded assessment, or (ideally) both. At the time of the examination of a participant, for example, an investigator may determine blood pressure two or more times and record the average. Performing the measurement without knowing the group assignment helps to minimize bias. In unblinded or single-blinded studies, the examination might be performed by someone other than the investigator, someone blind to the assignment. In assessing slides, X-rays, images, or electrocardiograms, two individuals can make independent, blinded evaluations, and the results can be averaged or adjudicated in cases of disagreement. Independent evaluations are particularly important when the assessment requires an element of judgment.

Centralized classification of major health outcomes by blinded reviewers is common in large clinical trials. The objective is to eliminate events that do not meet the protocol definitions. The process is thought to reduce the variability induced by having a large number of local investigators classifying fatal and major nonfatal events. The experience has been that a modest number of events are re-classified. A critical factor is how well the diagnostic criteria in a trial are specified and communicated to local investigators responsible for the initial classification. A recent review [28] based on the classification experience in 10 trials with over 9,000 cardiovascular events failed to detect any meaningful differences between initial classification and adjudication. It is unclear whether this observation also applies to other disease areas. The review raises questions about the validity of posthoc reclassifications of events that have reported major reversal of initial findings.

Data Entry

The introduction of computers into clinical trials has markedly improved data quality. Multicenter trials make increasing use of web-based functions. Systems have been developed for data entry but they have also been extended to include validation of forms and data, document management, tracking of shipments and specimens, queries and their resolution, scheduling, and adjudication processes [29]. Litchfield and coworkers [30] compared the efficiency and ease of use of internet data capture with the conventional paper-based data collection system. They reported substantial reductions with the internet-driven approach in terms of

time from visit to data entry, time to database release after the last participant visit, and time from a visit to a query being resolved. Seventy-one percent of the sites preferred the web-based approach. Different web-based systems have been developed. Examples include the Validation Studies Information Management System (VSIMS) [31], one developed for the Childhood Asthma Research (CARE) Network [32], and the Query and Notification System [33].

An issue under debate is whether paper forms can be totally eliminated. Often, a paper form is completed and the data transferred to a computer. Thus, a paper record trail is available for data verification and audit. The same record trail is essential for data entered directly into a computer. Programs have been developed which ensure that both original and revised data are saved, allowing an investigator to dispense with paper forms.

Quality Monitoring

Even though every effort is made to obtain high quality data, a monitoring or surveillance system is crucial. When errors are found, a monitoring system enables the investigator to take corrective action. Monitoring is most effective when it is current. Additionally, monitoring allows an assessment of data quality when interpreting study results. Numerous forms and procedures, including drug handling, can be monitored, but monitoring all of them is usually not feasible. Rather, monitoring those areas most important to the trial is recommended.

Monitoring of data quality proves most valuable when there is feedback to the clinic staff and technicians. Once weaknesses and errors have been identified, performance can be improved. Chapter 20 contains several tables illustrating quality control reports. With careful planning, reports can be provided and improvement can be accomplished without unblinding the staff. All quality control measures take time and money; it is thus difficult to be compulsive about the quality of every piece of data and every procedure. Investigators need to focus their efforts on those procedures which yield key data; those on which the conclusions of the study critically depend.

For clinical trials that will form the basis for regulatory decisions, the volume of data is very high and the data monitoring is very elaborate. Eisenstein and coworkers [34, 35] looked into ways of reducing the cost of large phase III trials. The major contributors to the expense are the number of case report form pages, the number of monitoring visits (for comparison of data in source records to the data on trial forms), and the administrative workload. Verification of critical information is important. Limiting the data verification of noncritical data may increase the error rate, but experience has shown that the overall rate is low and the effect on data quality limited. The cost of “queries” to resolve discrepancies can be very costly with estimates of more than $100 each. In sensitivity analyses, the authors showed that the total trial cost could be cut by more than 40% by reducing excessive

data collection and verification. Regular site visits to confirm that all case report forms are consistent with patient records seem excessive. As discussed below, sampling or selective site monitoring would be more appropriate in most situations.

Monitoring of Forms

During the study, key forms can be centrally checked electronically for completeness, internal consistency and consistency with other forms. When the forms disagree, the person or group responsible for ensuring consistent and accurate forms should question the person filling out the forms. Consistency within a given form can also be easily evaluated. Dates and times are particularly prone to error.

It may be important to examine consistency of data over time. A participant with a missing leg on one examination was reported to have palpable pedal pulses on a subsequent examination. Cataracts which did not allow for a valid eye examination at one visit were not present at the next visit, without an interval surgery having been performed. The data forms may indicate extreme changes in body weight from one visit to the next. In such a case, changing the data after the fact is likely to be inappropriate because the correct weights may be unknown. The observed differences in measurements may be less dramatic and not obvious. A quality control program based on randomly selected duplicate assessments has been advocated by Lachin [36]. However, the investigator can take corrective action for future visits by more carefully training his staff. Sometimes, mistakes can be corrected. In one trial, comparison of successive electrocardiographic readings disclosed gross discrepancies in the coding of abnormalities. The investigator discovered that one of the technicians responsible for coding the electrocardiograms was fabricating his readings. In this instance, correcting the data was possible.

Someone needs constantly to monitor completed forms to find evidence of missing participant visits or visits that are off schedule in order to correct any problems. Frequency of missing or late visits may be associated with the intervention. Differences between groups in missed visits may bias the study results. To improve data quality, it may be necessary to observe actual clinic procedures. Observing clinic procedures is particularly important in multicenter trials.

Monitoring of Procedures

Extreme laboratory values should be checked. Values incompatible with life such as, potassium of 10 mEq/l are obviously incorrect. Other, less extreme values (i.e., total cholesterol of 125 mg/dl in male adults in the United States who are not taking lipid-lowering agents) should be questioned. They may be correct, but it is unlikely. Finally, values should be compared with previous ones from the same participant. Certain levels of variability are expected, but when these levels are exceeded, the

value should be flagged as a potential outlier. For example, unless the study involves administering a lipid-lowering therapy, any determination which shows a change in serum cholesterol of perhaps 20% or more from one visit to the next should be repeated. Repetition would require saving samples of serum until the analysis has been checked. In addition to checking results, a helpful procedure is to monitor submission of laboratory specimens to ensure that missing data are kept to a minimum.

Investigators doing special procedures (laboratory work, electrocardiogram reading) need to have an internal quality control system. Such a system should include re-analysis of duplicate specimens or materials at different times in a blinded fashion. A system of resubmitting specimens from outside the laboratory or reading center might also be instituted. These specimens need to be indistinguishable from actual study specimens. An external laboratory quality control program established in the planning phase of a trial, can pick up errors at many stages (specimen collection, preparation, transportation, and reporting of results), not just at the analysis stage. Thus, it provides an overall estimate of quality. Unfortunately, the system most often cannot indicate at which step in the process errors may have occurred.

All recording equipment should be checked periodically. Even though initially calibrated, machines can break down or require adjustment. Scales can be checked by means of standard weights. Factors such as linearity, frequency response, paper speed, and time constant should be checked on electrocardiographic machines. In one long-term trial, the prevalence of specific electrocardiographic abnormalities was monitored. The sudden appearance of a threefold increase in one abnormality, without any obvious medical cause, led the investigator correctly to suspect electrocardiographic machine malfunction.

Monitoring of Drug Handling

In a drug study, the quality of the drug preparations should be monitored throughout the trial. Monitoring includes periodically examining containers for possible mislabeling and for proper contents (both quality and quantity). It has been reported that in one trial, “half of the study group received the wrong medication” due to errors at the pharmacy. Investigators should carefully look for discoloration and breaking or crumbling of capsules or tablets. When the agents are being prepared in several batches, samples from each batch should be examined and analyzed. Occasionally, monitoring the number of pills or capsules per bottle is useful. The actual bottle content of pills should not vary by more than 1 or 2%. The number of pills in a bottle is important to know because pill count may be used to measure adherence by participants.

Another aspect to consider is the storage shelf life of the preparations and whether they deteriorate over time. Even if they retain their potency, do changes in odor (as with aspirin) or color occur? If shelf life is long, preparing all agents at one time will minimize variability. Of course, in the event that the study ends prematurely, there may be a large supply of unusable drugs. Products having a short shelf

life require frequent production of small batches. Complete records should be maintained for all drugs prepared, examined, and used. Ideally, a sample from each batch should be saved. After the study is over, questions about drug identity or purity may arise and samples will be useful.

The dispensing of medication should also be monitored. Checking has two aspects. First, were the proper drugs sent from the pharmacy or pharmaceutical company to the clinic? If the study is double-blind, the clinic staff will be unable to check on this. They must assume that the medication has been properly coded. However, in unblinded studies, staff should check to assure that the proper drugs and dosage strengths have been received. In one case, the wrong strength of potassium chloride was sent to the clinic. The clinic personnel failed to notice the error. An alert participant to whom the drug was issued brought the mistake to the attention of the investigator. Had the participant been less alert, serious consequences could have arisen. An investigator has the obligation to be as careful about dispensing drugs as is a licensed pharmacist. Close reading of labels is essential, as well as documentation of all drugs that are handed out to participants.

Second, when the study is blinded, the clinic personnel need to be absolutely sure that the code number on the container is the proper one. Labels and drugs should be identical except for the code; therefore, extra care is essential. If bottles of coded medication are lined up on a shelf, it is relatively easy to pick up the wrong bottle accidentally. Unless the participant notices the different code, such errors may not be recognized. Even if she is observant, she may assume that she was meant to receive a different code number. The clinic staff should be asked to note on a study form the code number of the bottle dispensed and the code number of bottles that are returned by the participant. Theoretically, that should enable investigators to spot errors. In the end, however, investigators must rely on the care and diligence of the staff person dispensing the drugs.

It may be worthwhile periodically to send study drug samples to a laboratory for analysis. Although the center responsible for packaging and labeling drugs should have a foolproof scheme, independent laboratory analysis serves as an additional check on the labeling process.

The drug manufacturer assigns lot, or batch, numbers to each batch of drugs that are prepared. If contamination or problems in preparation are detected, then only those drugs from the problem batch need to be recalled. The use of batch numbers is especially important in clinical trials since the recall of all drugs can severely delay, or even ruin, the study. When only some drugs are recalled, the study can usually manage to continue. Therefore, the lot number of the drug as well as the name or code number should be listed in the participant's study record.

Audits

There are three general types of audits – routine audits of a random sample of records, structured audits, and audits for cause. Site visits are commonly conducted

in long-term multicenter trials. In many non-industry-sponsored trials, a 5–10% random sample of study forms may be audited for the purpose of verifying accurate transfer of data from hospital records. More complete audits are usually performed in industry-sponsored trials. Study monitors visit the sites in order to verify that the entered data are correct.

Some investigators have objections to random external data audits, especially in the absence of evidence of scientific misconduct. However, the magnitude of the problems detected when audits occur makes it difficult to take a position against them. Of interest, the FDA does not perform audits of trials sponsored by the National Cancer Institute (NCI) according to a long-standing agreement. It relies on a NCI-sponsored audit program that has been in place since 1982. A review of four cycles of internal audits conducted over a 11-year period by the investigators of the Cancer and Leukemia Group B (CLGB) showed similarities with FDA audits [37]. The deficiency rate (among main institutions) of 28% in the first cycle dropped to 13% in the fourth cycle. Only two cases of major scientific impropriety were uncovered during these on-site peer reviews. Compliance with institutional review board requirements improved over time, as did compliance with having proper consent forms. The consent form deficiencies dropped from 18.5% in the first cycle to 4% in the fourth. Although compliance with eligibility improved from 90 to 94%, no changes were noted for disagreement with auditors for treatment responses (5%) and deviations from the treatment protocol (11%). The authors concluded that the audit program had been successful in "pressuring group members to improve adherence to administrative requirements, protocol compliance, and data submission. It has also served to weed out poorly performing institutions."

Each of the 11 NCI cooperative groups has clearly established procedures for quality assurance. A detailed description of the CLGB system has been published [38]. NCI guidelines for monitoring are available online (http://ctep.cancer.gov/branches/ctmb/clinicalTrials/monitoring_coop_ccop_ctsu.htm). Another cooperative group, the National Surgical Adjuvant Breast and Bowel Project, conducted a review of almost 6,000 participant records [39]. The objective was to confirm participant eligibility, disease, and vital status. No additional treatment failures or deaths and only seven cases of ineligible participants were found. The audit was time-consuming and costly and since few discrepancies were found, the authors concluded that routine use of cooperative chart reviews cannot be supported. A similar conclusion was reached in the GUSTO trial [35]. Following an audit of all CRFs, the auditors reported only a small percentage of errors and that these errors did not change the trial conclusions.

The third type of audit is for cause, i.e., to respond to allegations of possible scientific misconduct. This could be expanded to include any unusual performance pattern such as enrolling participants well in excess of the number contracted for or anticipated. The Office of Research Integrity in the U.S. Department of Health and Human Services promotes integrity in biomedical and behavioral research sponsored by the U.S. Public Health Service at about 4,000 institutions worldwide. It monitors institutional investigations of research misconduct which includes fabrication,

falsification or plagiarism in proposing, performing, or reviewing research or in reporting research findings. In a review of 136 investigations resulting in scientific misconduct between 1992 and 2002, only 17 involved clinical trials. The most severe penalty, debarment from U.S. Government funding, was applied in six of the cases. Junior employees were often cited and the applied sanction was often a requirement for a plan of supervision to be implemented [40, 41].

The FDA conducts periodic audits as well as investigations into allegations of violations of the Federal Food, Drug, and Cosmetic Act through its Office of Criminal Investigations. These may include clinical investigator fraud such as falsifying documentation and enrolling ineligible patients. There were 2,866 FDA site inspections between 2000 and 2008. Most did not justify regulatory action and any corrective action was left to the investigator. Objectionable conditions were found and FDA sanctions indicated in 91 cases [42].

The quality of any trial is determined by the quality of its data. Experience has shown that too much data are being collected, much of which are never used for publication or review. As emphasized above, the data collection should be closely linked to the trial objectives and the questions posed in the protocol. Overcollection adds to the cost and effort of conducting the trial. Overemphasis on detailed audits of case report forms has similar effects. Moreover, the error rates are often so low that the value of most audits has been questioned. Rather, we should focus our quality control and auditing efforts on key variables. For other variables, samples should be audited with more reliance on statistical quality control procedures. Data collection in clinical trials should be streamlined whenever possible.

References

1. ICH E6. Good clinical practice: Consolidated guideline, Step 5 as of May 1996. http://www.ich.org.
2. International Conference on Harmonisation. September 2007. http://www.ich.org.
3. Knatterud GL, Rockhold FW, George SL, et al. Guidelines for quality assurance in multicenter trials: A position paper. *Control Clin Trials* 1998;19:477–493.
4. Williams GW. The other side of clinical trial monitoring; assuring data quality and procedural adherence. *Clin Trials* 2006;3:530–537.
5. Zon R, Meropol NJ, Catalano RB, Schilsky RL. American Society of Clinical Oncology statement on minimum standards and exemplary attributes of clinical trial sites. *J Clin Oncol* 2008;26:2562–2567.
6. Devine S, Dagher RN, Weiss KD, Santana VM. Good clinical practice and the conduct of clinical studies in pediatric oncology. *Pediatr Clin North Am* 2008;55:187–208.
7. Baigent C, Harrell FE, Buyse M, et al. Ensuring trial validity by data quality assurance and diversification of monitoring methods. *Clin Trials* 2008;5:49–55.
8. Acosta CJ, Galindo CM, Ochiai RL, et al. Implementation of good clinical practice guidelines in vaccine trials in developing countries. *Vaccine* 2007;25:2852–2857.
9. McFadden E. *Management of Data in Clinical Trials.* Hoboken, New Jersey: John Wiley & Sons, 2007.
10. Meinert CL. *Clinical Trials: Design, Conduct, and Analysis.* New York: Oxford University Press, 1986.

11. Neaton JD, Bartsch GE, Broste SK, et al. A case of data alteration in the Multiple Risk Factor Intervention Trial (MRFIT). *Control Clin Trials* 1991;12:731–740.
12. Fisher B, Redmond CK. Fraud in breast-cancer trials. *N Engl J Med* 1994;330:1458–1462.
13. Koran LM. The reliability of clinical methods, data and judgments. Part 1. *N Engl J Med* 1975;293:642–646.
14. Koran LM. The reliability of clinical methods, data and judgments. Part 2. *N Engl J Med* 1975;293:695–701.
15. Belk WP, Sunderman FW. A survey of the accuracy of chemical analyses in clinical laboratories. *Am J Clin Pathol* 1947;17:853–861.
16. Westgard JO. *Basic Method Validation.* Madison, Wisconsin: Westgard QC, Inc., 2003, pp 102–103.
17. McPherson RA, Pincus MR (eds.). *Henry's Clinical Diagnosis and Management by Laboratory Methods* (21st Edition). Philadelphia: Elsevier Saunders, Inc., 2007, pp 4–5.
18. Gur D, Bandos AI, Cohen CS, et al. The "laboratory" effect: Comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 2008;249:47–53.
19. Panicker GK, Karnad DR, Natekar M, et al. Intra- and interreader variability in QT interval measurement by tangent and threshold methods in central electrocardiogram laboratory. *J Electrocardiol* 2009;42:348–352.
20. Fridsma DB, Evans J, Hastak S, Mead CN. The BRIDG project: A technical report. *J Am Med Inform Assoc* 2008;15:130–137.
21. Weng C, Gennari JH, Fridsma DB. User-centered semantic harmonization: A case study. *J Biomed Inform* 2007;40:353–364.
22. Moyers S, Richesson R, Krischer J. Trans-atlantic data harmonization in the classification of medicines and dietary supplements: A challenge for epidemiologic study and clinical research. *Int J Med Inform* 2008;77:58–67.
23. Luepker RV, Evans A, McKeigue P, Reddy KS. *Cardiovascular Survey Methods* (3rd Edition). Geneva, Switzerland: World Health Organization, 2004.
24. Cook TD, DeMets DL. Data collection and quality control. In Cook TD, DeMets DL (eds.). *Introduction to Statistical Methods for Clinical Trials*. Boca Raton, Florida: Chapman & Hall/CRC, 2007, pp 171–200.
25. Kahn HA, Leibowitz H, Gauley JP, et al. Standardizing diagnostic procedures. *Am J Ophthalmol* 1975;79:768–775.
26. Neaton JD, Duchene AG, Svendson KH, Wentworth D. An examination of the efficacy of some quality assurance methods commonly employed in clinical trials. *Stat Med* 1990;9:115–124.
27. Furberg CD, Byington RP, Craven TE. Lessons learned from clinical trials with ultrasound endpoints. *J Intern Med* 1994;236:575–580.
28. Pogue J, Walter SD, Yusuf S. Evaluating the benefit of event adjudication of cardiovascular outcomes in large simple RCTs. *Clin Trials* 2009;6:239–251.
29. Reboussin D, Espeland MA. The science of web-based clinical trial management. *Clin Trials* 2005;2:1–2.
30. Litchfield J, Freeman J, Schou H, et al. Is the future for clinical trials internet-based? A cluster randomized clinical trial. *Clin Trials* 2005;2:72–79.
31. Winget M, Kincaid H, Lin P, et al. A web-based system for managing and co-ordinating multiple multisite studies. *Clin Trials* 2005;2:42–49.
32. Schmidt JR, Vignati AJ, Pogash RM, et al. Web-based distributed data management in the Childhood Asthma Research and Education (CARE) Network. *Clin Trials* 2005;2:50–60.
33. Mitchell R, Shah M, Ahmad S, et al. for the Adolescent Medicine Trials Network for HIV/AIDS interventions. A unified web-based query and notification system (QNS) for subject management, adverse events, regulatory, and IRB components of clinical trials. *Clin Trials* 2005;2:61–71.
34. Eisenstein EL, Lemons II PW, Tardiff BE, et al. Reducing the costs of phase III cardiovascular clinical trials. *Am Heart J* 2005;149:482–488.

35. Eisenstein EL, Collins R, Cracknell BS, et al. Sensible approaches for reducing clinical trial costs. *Clin Trials* 2008;5:75–84.
36. Lachin JM. The role of measurement reliability in clinical trials. *Clin Trials* 2004;1:553–566.
37. Weiss RB, Vogelzang NJ, Peterson BA, et al. A successful system of scientific data audits for clinical trials. *JAMA* 1993;270:459–464.
38. Weiss RB. Systems of protocol review, quality assurance, and data audit. *Cancer Chemother Pharmacol* 1998;42(suppl):S88–S92.
39. Soran A, Nesbitt L, Mamounas EP, et al. Centralized medical monitoring in phase III clinical trials: The National Surgical Adjuvant Breast and Bowel Project (NSABP) experience. *Clin Trials* 2006;3:478–485.
40. Reynolds SM. ORI findings of scientific misconduct in clinical trials and publicly funded research, 1992-2002. *Clin Trials* 2004;1:509–516.
41. Department of Health and Human Services. Public Health Service policies on research misconduct. Final rule. *Fed Regist* 2005;70:28370–28400.
42. Karlberg JPE. US FDA site inspection findings, 1997-2008, fail to justify globalization concerns. *Clin Trial Magnifier* 2009;2:194–212.

Chapter 12
Assessing and Reporting Adverse Events

There is no perfectly safe intervention. All treatments result in some adverse events. Their severity ranges from mild symptoms to life-threatening events. Collection of adverse event data in randomized clinical trials is a regulatory requirement and additionally, clinically and scientifically important. The challenge is to know what and how to collect these data, the frequency of collection, and how to deal with small numbers of serious events. There are also potential legal issues to consider, which tend to lead to an over-collection of safety data. On the other hand, there is a marked underreporting of safety information in the published literature. A review of 192 large clinical trials from seven therapeutic areas revealed that the safety reporting was considered adequate in only 39% of the articles [1].

The assessment of adverse events encompasses the whole spectrum of research from laboratory work during drug and device development, animal studies, and early work in small numbers of human beings, to case reports, clinical trials, and postmarketing surveillance. Carcinogenic or teratogenic consequences of drugs, such as noted with diethylstilbestrol [2–4] and thalidomide [5], thromboembolic events with COX-2 inhibitors [6–9], and suicides with antidepressants [10] or failures of devices such as cardiac pacemakers or silicone breast implants, received considerable publicity, but other sorts of findings are undoubtedly more common.

The literature related to assessing and reporting adverse events of devices is very limited. We see no reason to believe that the scientific challenges for devices are fewer or smaller than those encountered in drug development. The discussion in this chapter focuses on adverse events in clinical trials of drugs beyond the initial stages of development and testing. That is, even though the drugs may not yet have regulatory approval or be marketed, they have undergone early evaluation in human beings and are ready for larger scale evaluation. For the purposes of this book, adverse events are defined as any clinical event, sign, symptom, or laboratory or other finding that goes in an unwanted direction regardless of whether it is considered treatment-related.

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_12, © Springer Science+Business Media, LLC 2010

Fundamental Point

Adequate attention needs to be paid to the assessment, analysis, and reporting of adverse events to permit valid assessment of potential risks of interventions.

Clinical Trials in the Assessment of Adverse Events

Although the dual goals of a randomized clinical trial are to determine the efficacy and safety of an intervention, the assessment is generally asymmetric. Much more emphasis is placed on finding out whether and to what extent an intervention is beneficial. The primary, and most secondary, endpoints are with very few exceptions measures of efficacy. There are limitations, mostly ethical, on conducting a trial with safety as a primary outcome. Thus, the scope of a trial is determined by the sample size needed to find or refute a prespecified benefit. Similar consideration is rarely given to the power needed to confirm or dismiss a potentially serious adverse event (SAE), if not already part of the primary outcomes, such as mortality. The safety outcomes in a clinical trial protocol are often not very specific or prespecified. In spite of those limitations, clinical trials are an important source of safety information both from a regulatory and a clinical perspective.

There are three general categories of adverse events: (1) SAEs, (2) general adverse events (AEs), and (3) AEs of special interest. SAEs are defined as those events that are (a) life threatening, (b) result in hospitalization, (c) are irreversible, persistent, or significant disability/incapacity, or (d) are a congenital anomaly/birth defect. The SAEs are required to be reported to regulatory agencies within a fixed time period (e.g., 7 days) of their occurrence. General AEs are those which patients or trial participants have complained about or physicians have observed. These may range from very mild and not of much consequence to severe. In general, there is a great deal of variation in AE reporting. Due to this, some trials have designated certain AEs to be of special interest since they may seriously affect the interpretation and applicability of any new intervention. For example, these include liver function test abnormalities or changes in QT interval on an electrocardiogram.

Strengths

There are three distinct advantages to adverse event assessment in clinical trials. First, the safety determination can be obtained prospectively, which allows proper hypothesis testing and adds substantial credibility. Posthoc observations, common in the safety area, are often difficult to interpret in terms of causation and therefore, often lead to controversy.

Second, clinical trials by definition have a proper and balanced control group, which allows fair comparisons between the study groups. Other study designs have

a dilemma when comparing users of a particular treatment to nonusers. There is no guarantee that the user and nonuser groups are comparable. There are reasons why some patients get a particular intervention while others do not. Observed group differences can be treatment-induced, due to differences in the composition and characteristics of the groups, or a combination thereof. Statistical adjustments can help but will never be able to fully control the differences between users and nonusers.

Third, most drug trials allow for blinding, which reduces potential biases in the collection and reporting of safety data (Chap. 7).

Limitations in Identification of SAEs

There are four potential limitations in relying on clinical trials for safety evaluation. First, the trial participants are a selected sample of people with a given condition. The selectiveness is defined by the scope of the trial inclusion and exclusion criteria and the effect of being a volunteer. In general, trial participants are healthier than nonparticipants with the same disease. In addition, certain population groups may be excluded, for example, women who are pregnant or breastfeeding. Trials conducted prior to regulatory agency approval of the drug are typically designed to give clear findings of benefit and therefore often exclude from participation those who are old, have other medical conditions, and/or are taking other medications. The absence of SAEs observed in low-risk participants in preapproval trials is no assurance that a drug is safe when it reaches the marketplace. An early survey showed that most FDA-approved drugs have a SAE detected after approval when there is more exposure to higher-risk patients and longer treatment exposure [11]. More recent high-profile cases are the COX-2 inhibitors [6–9] and rosiglitazone [12–15].

A second limitation relates to the statistical power of finding a safety problem if it exists. Contributors to the limited power issue due to low SAE rates are small sample sizes and short trial durations as well as the focus on low-risk populations. Drug manufacturers may conduct a large number of small, short-term trials rather than fewer but larger trials of longer duration. Due to limited statistical power, clinical trials are unreliable for the detection of *rare* SAEs. Approximately 3,000 participants are required to detect a single case with 95% probability if the true incidence is one in 1,000; a total of 6,500 participants are needed to detect three cases [16]. When a new drug is approved for marketing, approximately 1,000–5,000 participants have typically been exposed to it. More commonly, the rare SAEs are initially discovered through case reports, other observational studies or reports of adverse events filed with regulatory agencies after approval [17, 18]. Vandenbroucke and Psaty [19] concluded that “the benefit side rests on data from randomized trials and the harms side on a mixture of randomized trials and observational evidence, often mainly the latter.”

Third, detection of *late* SAEs is another potential limitation of clinical trials. When a new compound or device is introduced, sometimes only several hundred

participants have been treated for 1 year or longer. This is obviously inadequate for evaluation of drugs intended for chronic or long-term use. Proper postmarketing studies are not always conducted to pursue safety signals noted in preapproval studies. A long lag-time to harm must be considered for drugs that may be carcinogenic or have adverse metabolic effects. The lag time for carcinogens to cause an increased incidence of cancer is generally longer than most long-term trials. We support the view that formal safety evaluation should continue the entire time a drug intended for chronic use is on the market [20].

Fourth, trials have a limited value for detecting *unexpected* SAEs. Data collection is typically decided prior to enrollment of the first trial participant. What is unexpected is by definition, not prespecified and is not included in data collection forms. The only way to collect information on unexpected SAEs is to have open-ended safety questions and alert investigators.

The optimal collection of safety information is to take advantage of the strengths of clinical trials and to supplement them with properly designed and conducted observational studies, especially if safety issues or signals emerge. Establishment of such long-term safety registries is becoming more common [21].

Determinants of Adverse Events

Definitions

The rationale for defining adverse events is similar to that for defining any response variable; it enables investigators to record something in a consistent manner. Further, it allows someone reviewing a trial to assess it better, and possibly to compare the results with those of other trials of similar interventions.

Because adverse events are typically viewed as secondary or tertiary response variables, they are not often seriously thought about ahead of time with the same degree of attention. Generally, an investigator will prepare a list of potential adverse events on a study form. They usually are not defined, except by the way investigators define them in their daily practice. Study protocols seldom contain written definitions of adverse events, except for those that are recognized clinical conditions. In multicenter trials, the situation may often be even worse. In those cases, an adverse event may be simply what each investigator declares it to be. Thus, intrastudy consistency may be as poor as interstudy consistency.

However, given the large number of possible adverse events, it is not feasible to define all of them in advance and many do not lend themselves to good definition. Some adverse events cannot be defined because they are not listed in advance, but are spontaneously mentioned by the participants. Though it is not always easy, important adverse events which are associated with individual signs or laboratory findings, or a constellation of related signs, symptoms, and

laboratory results can and should be well-defined. These include ones known to be associated with the intervention and which are clinically important, i.e., AEs of special interest. Other adverse events that are purely based on a participant's report of symptoms may be important but are more difficult to define. These may include nausea, fatigue, or headache. Changes in the degree of severity of any symptom would also meet the definition of an adverse event. The fact that an adverse event is not well-defined or not prespecified should be stated in any trial publication.

Classification of Adverse Events

A major step toward the development of common international medical terminology for adverse events was taken in 1994 by the International Conference on Harmonisation (ICH). This stemmed from a need from the biopharmaceutical establishment to standardize regulatory communication across countries or regulatory jurisdictions. Version 2 of the Medical Directory for Regulatory Activities (MedDRA) Terminology was introduced in 1997 as one such system [22]. Included are categories of terms for signs, symptoms, diseases, diagnoses, procedures, and others. The terms are structured hierarchically. Lowest Level Terms (LLTs) include synonyms and provides maximum specificity. Preferred Terms (PTs) constitute single medical concepts. The latest edition included more than 66,000 LLTs and more than 18,000 PTs. The highest level of the hierarchy is the System Organ Class (SOC), of which there are 26. A challenge is that the LLT terms are so granular that it is difficult to identify a real signal. As a result, individual items are not frequent enough to detect statistically different event rates. On the other hand, the higher order terms contain important adverse events, but these are mixed with less important ones and noise.

MedDRA is an internationally accepted system for classification of adverse events. It is a commercial system only available by subscription and, as a result, it may perhaps not be affordable to all potential users. New editions are released semiannually [22]. This updating introduces its own problems in clinical trials of a duration more than a few months. Recoding of early trial data may be required in trials of longer duration. The strength of MedDRA is the ease of use for data entry, retrieval, analysis, and display.

The National Cancer Institute (NCI) Common Terminology Criteria for Adverse Events v3.0 is another advanced system for reporting adverse events (http://ctep.cancer.gov). It is structured with a broad classification of adverse events based on anatomy and pathophysiology. There are 28 categories of adverse events and within each are a large number of specific events. Another strength is the five-step severity scale for each adverse event ranging from mild (grade 1) to any fatal adverse event (grade 5). It is free of charge.

Ascertainment

The issue of whether one should elicit adverse events by means of a checklist or rely on the participant to volunteer complaints often arises. Eliciting adverse events has the advantage of allowing a standard way of obtaining information on a preselected list of symptoms. Thus, both within and between trials, the same series of events can be ascertained in the same way, with assurance that a "yes" or "no" answer will be present for each. This presupposes, of course, adequate training in the administration of the questions. Volunteered responses to a question such as "Have you had any health problems since your last visit?" have the possible advantage of tending to yield only the more serious episodes, while others are likely to be ignored or forgotten. In addition, only volunteered responses will give information on truly unexpected adverse events.

In the Aspirin Myocardial Infarction Study [23], information on several adverse events was both volunteered by the participants and elicited. After a general question about adverse events, the investigators asked about specific complaints. The results for three adverse events are presented in Table 12.1. Two points might be noted. First, for each adverse event, eliciting gave a higher percent of participants with complaints than did asking for volunteered problems. Second, the same aspirin–placebo differences were noted, regardless of the method. Thus, the investigators could arrive at the same conclusions with each technique. In this study, little additional information was gained by the double ascertainment. Perhaps the range between the volunteered and the solicited numbers within the individual study groups provides bounds on the true incidence of the adverse event.

Downing et al. reported on a comparison of elicited versus volunteered adverse events in a trial of tranquilizers and antidepressants [24]. Thirty-three participants on active drug volunteered complaints, as opposed to 12 on placebo. This contrasts with 53 elicited complaints from the active drug group and 12 elicited from the placebo group. The authors concluded that eliciting adverse events preferentially increases the number in the active drug group, rather than the placebo group. This is contrary to the findings in the Aspirin Myocardial Infarction Study [23]. Of 29 drug-treated participants who had complaints ascertained by both eliciting and volunteering, 26 were classified as more severe. Of 24 participants whose complaints were ascertained only by eliciting, half were called more severe. Therefore, the requirement that an adverse event be volunteered by a participant

Table 12.1 Percent of participants ever reporting (volunteered and solicited) selected adverse events, by study group, in the Aspirin Myocardial Infarction Study

	Hematemesis	Tarry stools	Bloody stools
Volunteered			
Aspirin	0.27	1.34[a]	1.29[b]
Placebo	0.09	0.67	0.45
Elicited			
Aspirin	0.62	2.81[a]	4.86[a]
Placebo	0.27	1.74	2.99

[a]Aspirin–placebo difference > 2 S.E
[b]Aspirin–placebo difference > 3 S.E

led to a preponderance of severe ones. Because of these and other inconsistent findings, many researchers have continued to use both methods.

It has been suggested that subjective adverse events are influenced by the amount of information provided to participants during the informed consent process. Romanowski and colleagues compared responses of 25 people given general information about possible adverse events in a consent form with responses from the 29 provided with a detailed listing of possible adverse events [25]. In this study, there was no important difference in frequency of reported subjective adverse events (4 vs. 6). The investigators therefore concluded that "previous priming of the patient" did not affect reporting of adverse events. Obviously, the numbers are small, and a larger study would be necessary to confirm this.

Dimensions

The simplest way of recording an adverse event is with a yes/no answer. This information is likely to be adequate if the adverse event is a serious clinical event such as a stroke, a hospitalization, or a significant laboratory abnormality. However, symptoms have other important dimensions such as severity and frequency of occurrence.

The severity of subjective symptoms is typically rated as mild, moderate, or severe. However, the clinical relevance of this rating is unclear. Participants have different thresholds for perceiving and reporting their reactions. In addition, staff's recorded rating of the reported symptom may also vary. One way of dealing with this dilemma is to consider the number of participants who were taken off the study medication due to the adverse event, the number who had their dose of the study medication reduced, and those who continued treatment according to protocol in spite of a reported adverse symptoms. This classification of severity makes clinical sense and is generally accepted. A challenge may be to decide how to deal with participants who temporarily are withdrawn from study medication or have their doses reduced.

The frequency with which a particular adverse event occurs in a participant can be viewed as another measure of severity. For example, episodes of nausea occurring daily rather than monthly, are obviously more troublesome to the participant. Presenting such data in a clear fashion is complicated.

Presence versus absence of an adverse event, severity, and frequency are dimensions of drug safety that need to be considered in the planning of a trial.

Length of Follow-Up

Obviously, the duration of a trial has a substantial impact on adverse event assessment. The longer the trial, the more opportunity one has to discover adverse events, especially those with low frequency. Also, the cumulative number of participants in the intervention group complaining will increase, giving a better estimate of the adverse event incidence. Of course, eventually, most participants will report some

general complaint, such as headache or fatigue. However, this will occur in the control group as well. Therefore, if a trial lasts for several years, and an adverse event is analyzed simply on the basis of cumulative number of participants suffering from it, the results may not be very informative.

Duration of follow-up is also important in that exposure time may be critical. Some drugs may not cause certain adverse events until a person has been taking them for a minimum period. An example is the lupus syndrome with procainamide. Given enough time, a large proportion of participants will develop this syndrome, but very few will do so if followed for only several weeks. Other sorts of time patterns may be important as well. Many adverse events even occur at low drug doses shortly after initiation of treatment. In such circumstance, it is useful, and indeed prudent, to monitor carefully participants for the first few hours or days. If no reactions occur, the participant may be presumed to be at a low risk of developing these events subsequently.

In the Diabetes Control and Complications Trial (DCCT) [26], cotton exudates were noted in the eyes of the participants receiving tight control of glucose early after onset of the intervention. Subsequently, the progression of retinopathy in the regular control group surpassed that in the tight control group, and tight control was shown to reduce retinal complications in insulin dependent diabetes. Focus on only this short-term adverse event might have led to early trial termination. Fortunately, DCCT continued and reported a favorable long-term risk-benefit balance.

Figure 12.1 illustrates the occurrence of ulcer symptoms and complaints of stomach pain, over time, in the Aspirin Myocardial Infarction Study. Ulcer symptoms

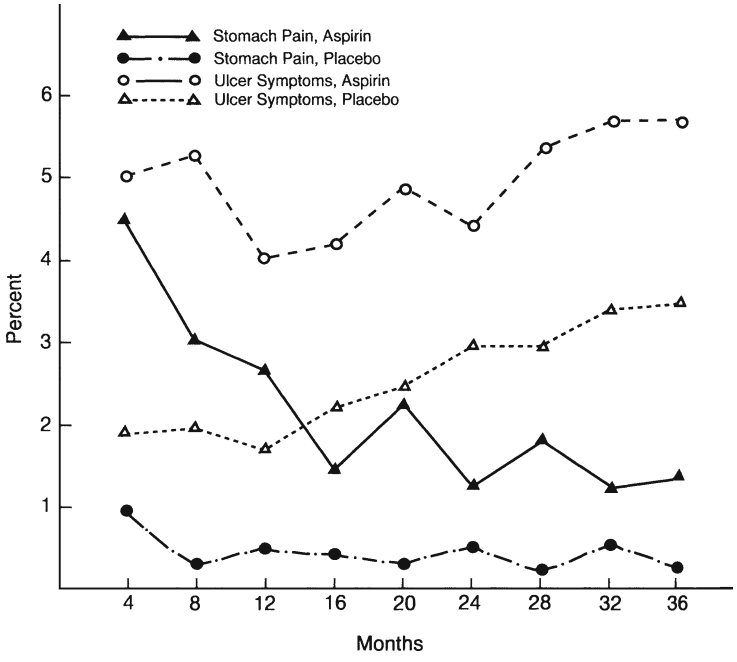


Fig. 12.1 Percent of participants reporting selected adverse events, over time, by study group, in the Aspirin Myocardial Infarction Study

rose fairly steadily in the placebo group, peaking at 36 months. In contrast, complaints of stomach pain were maximal early in the aspirin group, and then they decreased. Participants on placebo had a constant, low level of stomach pain complaints. If a researcher tried to compare adverse events in two studies of aspirin, one lasting weeks and the other months, his findings would be different. To add to the complexity, the aspirin data in the study of longer duration may be confounded by changes in aspirin dosage and concomitant therapy.

An intervention may cause continued discomfort throughout a trial, and its persistence may be an important feature. Yet, unless the discomfort is considerable, such that the intervention is stopped, the participant may eventually stop complaining about it. Unless the investigator is alert to this possibility, the proportion of participants with symptoms at the final assessment in a long-term trial may be misleadingly low.

Analyzing Adverse Events

Types of Analysis

Analyzing the presence versus absence of an adverse event and the proportion of participants withdrawn from treatment due to adverse events or having their dose reduced versus continuing the study intervention is straightforward. The challenge is how to characterize participants who have their treatment temporarily stopped or reduced. These occurrences should be noted. One approach is to present these data by visit.

The fact that the number of adverse event types can be substantial raises the issue of multiple testing (Chap. 17). Because of this problem, the intervention may appear significantly different from the control more often by chance alone than indicated by the *p*-value. Nevertheless, investigators tend to relax their statistical requirements for declaring adverse events to be real findings. It reflects understandable conservatism and the desire to avoid unnecessary harm to the participants. However, we should always keep in mind that this conclusion might be incorrect.

As with other response variables, adverse events can be analyzed using survival analysis methods (Chap. 15). An advantage of this sort of presentation is that the time to a particular episode, in relation to when the intervention was started, is examined. Further, the frequency of a particular adverse event will be directly related to the number of participants at risk of suffering it. This can give a higher rate of adverse events than other measures, but this high rate may be a realistic estimate. Difficulties with survival analysis techniques include the problems of not considering repeated episodes in any participant (i.e., only the first episode is counted) and severity of a particular adverse event, changes in dosing pattern or adherence, and changes in sensitization or tolerance to adverse events. Nevertheless, this technique has been underutilized in reporting adverse events.

Analysis of Data from Nonadherent Participants

As discussed in Chap. 17, there are differing views on analyzing data from participants who fail to adhere to the study intervention regimen. For analysis of primary response variables in the typical superiority trial, the "intention-to-treat" approach, which includes all participants in their originally randomized groups, is more conservative and less open to bias than the "explanatory" approach, which omits participants who stop taking their assigned intervention. When adverse events are assessed, however, the issue is less clear. Participants are less likely to report adverse events if they are off medication (active or placebo) than if they are on it. Therefore, analyzing event rates by level of adherence may underestimate their true incidence. The explanatory approach makes it impossible to assess events which occur sometime after drug discontinuation, but may in fact be a real adverse event that was not recognized until later. Participants are sometimes followed for a short period of time (7–30 days) after discontinuation of the medication to allow it to be "washed out." However, as illustrated in the APPROVe study (see Chap. 17), as excess of treatment-induced major events were observed during the first year after treatment was stopped. In addition, withdrawing participants can void the benefits of randomization, resulting in invalid group comparisons. While there is no easy solution to this dilemma, it is probably safer and more reasonable to continue to assess adverse events for the duration of the trial, even if a participant has stopped taking his study drug. The analysis and reporting might then be done both including and omitting nonadherent participants [27]. Certainly, it is extremely important to specify what was done. In conclusion, the intention-to-treat approach is the preferred one.

Reporting of Adverse Events

Scientific

The usual measures of adverse events include the following:

(a) Reasons participants are taken off study medication or device removed.
(b) Reasons participants are on reduced dosage of study medication or on lower intensity of intervention.
(c) Type and frequency of participant complaints.
(d) Laboratory measurements, including X-rays and imaging.
(e) In long-term studies, possible intervention-related reasons participants are hospitalized.
(f) Combinations or variations of any of the above.

All of these can rather easily indicate the number of participants with a particular adverse event during the course of the trial. Presenting the frequency of adverse events in a clear fashion is complicated. It can be done by means of frequency

Table 12.2 Percent of participants with drug dosage reduced or complaining of selected adverse events, by study group, in the Aspirin Myocardial Infarction Study

	Aspirin ($N = 2,267$)	Placebo ($N = 2,257$)
Hematemesis		
Reasons dosage reduced	0.00	0.00
Complaints	0.27	0.09
Tarry stools		
Reasons dosage reduced	0.09	0.04
Complaints	1.34	0.67
Bloody stools		
Reasons dosage reduced	0.22	0.04
Complaints	1.29	0.45

distributions, but these consume considerable space in tables. Another method is to select a frequency and assume that adverse events which occur less often in a given time period are less important. Thus, only the number of participants with a frequency of specified adverse events above that are reported. As an example, of ten participants having nausea, three might have it at least twice a week, three at least once a week, but less than twice, and four less than once a week. Only those six having nausea at least once a week might be included in a table. These ways of reporting assume that adequate and complete data have been collected, and may require the use of a diary. Obviously, if a follow-up questionnaire asks only if nausea has occurred since the previous evaluation, frequency measures cannot be presented.

Severity indices can be more complicated. It may be assumed that a participant who was taken off study drug because of an adverse event had a more serious episode than one who merely had his dosage reduced. Someone who required dose reduction probably had a more serious event than one who complained but continued to take the dose required by the study protocol. Data from the Aspirin Myocardial Infarction Study [23], using the same adverse events as in the previous example, are shown in Table 12.2. In the aspirin and placebo groups, the percents of participants complaining about hematemesis, tarry stools, and bloody stools are compared with the percents having their medication dosage reduced for those events. As expected, numbers complaining were many times greater than the numbers with reduced dosage. Thus, the implication is that most of the complaints were for relatively minor occurrences or had been transient.

As previously mentioned, another way of reporting severity is to establish a hierarchy of consequences of adverse events, such as permanently off-study drug, which is more severe than permanently on reduced dosage, which is more severe than ever on reduced dosage, which is more severe than ever complaining about the event. Unfortunately, few clinical trial reports present such severity data.

Published Reports

Published reports of clinical trials typically emphasize the favorable results. The harmful effects attributed to a new intervention are often incompletely reported.

This discordance undermines an assessment of the risk-benefit balance. A review of randomized clinical trials published in 1997 and 1998 showed that safety reporting varied widely and, in general, was inadequate [1]. Several subsequent studies evaluating the reporting of harm in clinical trials came to the same conclusion [28].

In an effort to improve this inadequate and troubling reporting of adverse events, the CONSORT statement added specific guidelines for reporting harm-related results of clinical trials in 2004 [29]. Included in the reporting should be descriptions of adverse events with numerical data by treatment group, information related to the severity of adverse events, and the number of participants withdrawn from their study medications due to adverse events.

In the first 2 years after the publication of the CONSORT guidelines, the impact was negligible. Pitrou et al [30] analyzed 133 reports of randomized clinical trials published in six general medical journals in 2006. No adverse events were reported in 11% of the reports. Eighteen percent did not provide numerical data by treatment group and 32% restricted the reporting to the most common events. The data on severity of adverse events were missing in 27% of the publications and almost half failed to report the proportion of participants withdrawn from study medication due to adverse events. It is imperative that investigators devote more attention to reporting the key safety data from their clinical trials in the main results article. Additional safety data could be included in appendices to this paper or covered in separate articles.

Regulatory

The major drug regulatory agencies in the world have a number of requirements for expedited reporting of adverse events [31, 32]. This requirement applies to serious, unexpected, and drug-related events. As described earlier, serious is defined as death, life-threatening experience, inpatient hospitalization or prolongation of hospitalization, persistent or significant disability/incapacity, or congenital anomaly/birth defect. The event must be reported in writing within 15 calendar days of being informed. For a death or life-threatening event, the report should be made by fax or telephone within 7 days of notification. The regulations do not specify deadlines for sites to report these events to the study sponsor, although sponsors typically establish their own deadlines.

The purpose of premarketing risk assessment is to identify adverse drug events prior to any regulatory approval for marketing [33–35]. This assessment is typically incomplete for several reasons. Very few phase 3 trials are designed to test specified hypotheses about safety. They are often too small to detect less common SAEs or AEs of special interest and additionally, the assessment of multiple adverse events raises questions regarding proper significance levels. Moreover, the premarketing trials tend to focus on low-risk patients by excluding elderly persons, those with other medical conditions, and those on concomitant medications, which also reduces the statistical power. This focus on low-risk patients leads to an underestimation of safety issues in future users of the medication.

To address these power problems, data from multiple trials are often combined. A concern is that meta-analyses of heterogeneous trials can obscure meaningful differences between trials. Adverse events are more likely to occur in trials testing higher doses of a new drug in long-term trials and in trials enrolling more vulnerable patients, i.e., elderly patients with multiple medical conditions who use multiple medications.

To deal with often limited safety information, special attention is given to trends in the data. A *safety signal* [34] is defined as "a concern about an excess of adverse events compared to what would be expected to be associated with a product's use." These signals generally indicate a need for further investigation in order to determine whether it is drug-induced or a chance finding.

Rules for reporting adverse events to the local institutional review boards (IRBs) vary. Many require that investigators report all events meeting regulatory agency definitions. The IRB has, based on the safety report, several options. These include making no change, requiring changes to the informed consent and the trial protocol, placing the trial on hold, or terminating approval of the trial. However, the IRBs seldom have the expertise or infrastructure to deal with serious or adverse event reports from multicenter trials, or even local trials. When the trial is multicenter, different rules and possible actions can cause considerable complications. These complications can be reduced when IRBs agree to rely on safety review by a study-wide monitoring committee.

Identification of SAEs

As pointed out earlier in this chapter, randomized clinical trials are not optimal for the detection of rare, late, and unexpected SAEs. Experience has shown that critical information on serious events comes from multiple sources.

The role of clinical trials in identifying SAEs was investigated in an early study by Venning [31], who reviewed the identification and report of 18 adverse events in a variety of drugs. Clinical trials played a key role in identifying only three of the 18 adverse events discussed. Of course, clinical trials may not have been conducted in all of the other instances. Nevertheless, it is clear that assessment of adverse events, historically, has not been a major contribution of clinical trials. As pointed out earlier in this chapter, observational studies conducted postmarketing contribute more to the identification of harmful drug effects than randomized trials [19]. A comparison of evidence of 15 harms of various interventions in large randomized and nonrandomized studies showed that the nonrandomized studies often were more conservative in the estimates of risk [36].

A clinical trial may, however, suggest that further research on adverse events would be worthwhile. As a result of implications from the Multiple Risk Factor Intervention Trial [37] that high doses of thiazide diuretics might increase the incidence of sudden cardiac death, Siscovick and colleagues conducted a population-based case-control study [38]. This study confirmed that high doses of thiazide diuretics, as opposed to low doses, were associated with a higher rate of cardiac arrest.

Drugs of the same class generally are expected to have similar events on the primary outcome of interest. For example, different angiotensin converting enzyme inhibitors will reduce blood pressure and ease symptoms of heart failure. Different calcium channel blocking agents will treat hypertension and angina. The factors that make the drugs in the same class different, however, may mean that adverse events may differ, in degree if not in kind. One illustration is cerivastatin which was much more likely to cause rhabdomyolysis than the other statins [39]. Longer acting preparations, or preparations that are absorbed or metabolized differently, may be administered in different doses and have greater or lesser adverse events. It cannot be assumed in the absence of appropriate comparisons that the adverse events from similar drugs are or are not alike. As noted, however, a clinical trial may not be the best vehicle for detecting these differences, unless it is sufficiently large and of long duration.

Potential Solutions

First, one obvious solution to the problem is the conduct of larger and longer clinical trials in participants who better represent future users of the medication. This is unlikely to happen unless the regulatory requirements for drug approval are changed and manufacturers are required to submit better safety data on their new products. Another question is what ethics and medical practice will allow for drugs already on the market.

Second, when individual trials are inconclusive, the fall-back position is the combination of safety data from multiple trials in a meta-analysis or systematic review (see Chap. 17). A major contributor in this area is the Cochrane Collaboration, which established the Cochrane Adverse Effects Methods Group to address the many challenges in systematic reviews of adverse events (http://aemg.cochrane.org/en/index.html). Its charge is to raise awareness of the adverse effects of interventions, to develop research methodologies, and to advise on how to improve validity and precision in systematic reviews of adverse effects. The group organizes seminars and workshops to train individuals on the advanced methodological techniques for conducting systematic reviews of adverse effects. A list of relevant publications from members of the group and other papers can be found at http://aemg.cochrane.org/en/publications.html. In one of the systematic reviews, Golder and Loke [40] concluded that industry funding may not be a major threat to biased reporting of adverse event data, but that there are concerns related to the interpretation and conclusions of these data.

Meta-analyses conducted by manufacturers are commonly included in New Drug Applications submitted to regulatory agencies. There is also an increasing number of meta-analyses of treatment safety published in leading medical journals. Singh and coworkers published three meta-analyses showing that rosiglitazone and pioglitazone double the risk of heart failure and fractures (in women) in type 2 diabetes [12, 14] and that rosiglitazone, in contrast to pioglitazone, also increases

the risk of heart attacks [13]. None of these adverse effects were recognized at the time of regulatory approval of these drugs. It has been recommended that cumulative meta-analysis be conducted to determine whether and when pooled safety data reveal increased risk [41]. The authors concluded that cumulative clinical trial data revealed increased cardiovascular risk associated with rofecoxib a couple of years before the drug was withdrawn from the U.S. market.

It is important to keep in mind that meta-analyses may be misleading. On one hand, individual trials revealing unfavorable results may never be reported or published. Thus, publication bias can lead to an underestimation of the true rate of adverse effects. On the other hand, small meta-analyses do not always provide accurate information. Experience has shown that conclusions from meta-analyses of a large number of small trials at times are not confirmed in subsequent large trials. An illustration of how event data can fluctuate when the numbers are small can be seen in Fig. 3 of Chap. 16. The mortality data came close early to a 0.05 significance level on 3–4 occasions, only to return to no difference at the trial completion [42]. Thus, caution is always advised when the numbers are small.

Third, the field of pharmacogenetics holds promise for better identification in the future of patient groups that are more likely to develop SAEs (see Chap. 9).

Fourth, observational studies will always have a role in the identification of SAEs. A detailed discussion of these types of studies falls outside the scope of this book. However, the use of very large observational studies has been successfully used in the past [36]. However, since observational studies rely on a comparison of users and nonusers of a particular treatment, their comparability is critical. Even extensive covariate adjustments cannot guarantee comparability between users and nonusers in a way that randomization does it for clinical trials.

Fifth, other potential solutions are case-control studies and databases. The former have similar limitations as other observational studies. The reliance on database studies depends on the accuracy and completeness of the recorded safety information. For example, data on suicides and suicidal ideation may not be included in the typical database. An attractive way of shortening the time from marketing of a product to the identification of new adverse events is employed in New Zealand [43]. The first 5–10,000 users of a newly approved drug are registered. After a given time, for example, after 6 months, they are all contacted and asked about their experiences with the drug. This approach could detect new safety signals early. A limitation is the lack of a comparable control group.

In the end, given the limitations of each approach, a combination of them will remain the most important way to identify and assess adverse events.

References

1. Ioannidis JPA, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. *JAMA* 2001;285:437–443.
2. Herbst AL, Ulfelder H, Poskanzer DC. Adenocarcinoma of the vagina. Association of maternal stilbestrol therapy with tumor appearance in young women. *N Engl J Med* 1971;284:878–881.

3. Heinonen OP, Slone D, Shapiro S. *Birth Defects and Drugs in Pregnancy.* Littleton: PSG Publishing Company, 1977, pp. 1–7.
4. Meador KJ, Baker GA, Browning N, et al. for the NEAD Study Group. Cognitive function at 3 years of age after fetal exposure to antiepileptic drugs. *N Engl J Med* 2009;360:1597–1605.
5. McBride WG: Thalidomide and congenital malformations. *Lancet* 1961;*ii*:1358.
6. Bombardier C, Laine L, Reicin A, et al. for the VIGOR Study Group. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. *N Engl J Med* 2000;343:1520–1528.
7. Bresalier RS, Sandler RS, Quan H, et al. for the Adenomatous Polyp Prevention on Vioxx (APPROVe) Trial Investigators. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. *N Engl J Med* 2005;352:1092–1102.
8. Solomon SD, McMurray JJV, Pfeffer MA, et al. for the Adenoma Prevention with Celecoxib (APC) Study Investigators. Cardiovascular risk associated with celecoxib in a clinical trial for colorectal adenoma prevention. *N Engl J Med* 2005;352:1071–1080.
9. Psaty BM, Furberg CD. COX-2 inhibitors – lessons in drug safety. *N Engl J Med* 2005;352:1133–1135.
10. Fergusson D, Doucette S, Glass KC, et al. Association between suicide attempts and selective serotonin reuptake inhibitors: systematic review of randomised controlled trials. *Br Med J* 2005;330: 396–399. doi:10.1136/bmj.330.7488.396 (published 19 February 2005).
11. US General Accounting Office. *FDA Drug Review: Postapproval Risks, 1976–85*. Washington, DC: US General Accounting Office, April 26, 1990, GAO/PEMD-90-15.
12. Singh S, Loke YK, Furberg CD. Thiazolidinediones and heart failure. *Diabetes Care* 2007;30:2141–2153.
13. Singh S, Loke YK, Furberg CD. Long-term risk of cardiovascular events with rosiglitazone: a meta-analysis. *JAMA* 2007;298:1189–1195.
14. Loke YK, Singh S, Furberg CD. Long-term use of thiazolidinediones and fractures in type 2 diabetes: a meta-analysis. *CMAJ* 2009;180:32–39.
15. Fong DS, Contreras R. Glitazone use associated with diabetic macular edema. *Am J Ophthalmol* 2009;147:583–586.
16. Furberg BD, Furberg CD. *Evaluating Clinical Research. All that Glitters is Not Gold* (2nd edition). New York, NY: Springer, 2007, pp. 17–18.
17. Venning GR. Identification of adverse reactions to new drugs. II: How were 18 important adverse reactions discovered and with what delays? *Br Med J* 1983;286:289–292 and 365–368.
18. Aronson JK, Derry S, Loke YK. Adverse drug reactions: keeping up to date. *Fundam Clin Pharmacol* 2002;16:49–56.
19. Vandenbroucke JP, Psaty BP. Benefits and risks of drug treatments. How to combine the best evidence on benefits with the best data about adverse effects. *JAMA* 2008;300:2417–2419.
20. Committee on the Assessment of the US Drug Safety System. Baciu A, Stratton K, Burke SP (eds.). *The Future of Drug Safety: Promoting and Protecting the Health of the Public.* Washington, DC: The National Academies Press, 2006.
21. Furberg CD, Levin AA, Gross PA, et al. The FDA and drug safety. A proposal for sweeping changes. *Arch Intern Med* 2006;166:1938–1942.
22. MedDRA and the MSSO. http://www.meddramsso.com/MSSOWeb/index.htm.
23. Aspirin Myocardial Infarction Study Research Group. A randomized, controlled trial of aspirin in persons recovered from myocardial infarction. *JAMA* 1980;243:661–669.
24. Downing RW, Rickels K, Meyers F. Side reactions in neurotics: 1. A comparison of two methods of assessment. *J Clin Pharmacol* 1970;10:289–297.
25. Romanowski B, Gourlie B, Gynp P. Biased adverse effects? *N Engl J Med* 1988;319:1157–1158.
26. The Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med* 1993;329:977–986.

27. Baron JA, Sandler RS, Bresalier RS, et al. Cardiovascular events associated with rofecoxib: final analysis of the APPROVe trial. *Lancet* 2008;372:1756–1764.
28. Ioannidis JPA. Adverse events in randomized trials. Neglected, restricted, distorted, and silenced. *Arch Intern Med* 2009;169:1737–1739.
29. Ioannidis JPA, Evans SJ, Gøtzsche PC, et al. for the CONSORT Group. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med* 2004;141:781–788.
30. Pitrou I, Boutron I, Ahmad N, Ravaud P. Reporting of safety results in published reports of randomized controlled trials. *Arch Intern Med* 2009;169:1756–1761.
31. Department of Health and Human Services, Food and Drug Administration: International Conference on Harmonisation; Guideline on clinical safety data management: Definitions and standards for expedited reporting, Notice. *Federal Register* 60 (1 March 1995):11284–11287.
32. Department of Health and Human Services, Food and Drug Administration. International Conference on Harmonisation; Draft guidance on E2D postapproval safety data management: Definitions and standards for expedited reporting, Notice. *Federal Register* 68 (15 September 2003):53983–53984.
33. U.S. Department of Health and Human Services. Food and Drug Administration. Guidance for Industry. Premarketing risk assessment. March 2005. www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm126958.pdf.
34. U.S. Department of Health and Human Services. Food and Drug Administration. Guidance for Industry. Good pharmacovigilance practices and pharmacoepidemiologic assessment. March 2005. http://www.fda.gov/downloads/RegulatoryInformation/Guidances/UCM126834.pdf.
35. U.S. Department of Health and Human Services. Food and Drug Administration. Reviewer Guidance. Conducting a clinical safety review of a new product application and preparing a report on the review. March 2005. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM072974.pdf.
36. Papanikolaou PN, Christidi GD, Ioannidis JPA. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. *CMAJ* 2006;174:635–641.
37. Multiple Risk Factor Intervention Trial Research Group. Baseline rest electrocardiographic abnormalities, antihypertensive treatment, and mortality in the Multiple Risk Factor Intervention Trial. *Am J Cardiol* 1985;55:1–15.
38. Siscovick DS, Raghunathan TE, Psaty BM, et al. Diuretic therapy for hypertension and the risk of primary cardiac arrest. *N Engl J Med* 1994;330:1852–1857.
39. Psaty BM, Furberg CD, Ray WA, Weiss NS. Potential for conflict of interest in the evaluation of suspected adverse drug reactions: use of cerivastatin and risk of rhabdomyolysis. *JAMA* 2004;292:2622–2631.
40. Golder S, Loke YK. Is there evidence for biased reporting of published adverse effects data in pharmaceutical industry-funded studies? *Br J Clin Pharmacol* 2008;66:767–773.
41. Ross JS, Madigan D, Hill KP, et al. Pooled analysis of rofecoxib placebo-controlled clinical trial data. Lessons for postmarket pharmaceutical safety surveillance. *Arch Intern Med* 2009;169:1976–1984.
42. Hennekens CH, DeMets D. The need for large-scale randomized evidence without undue emphasis on small trials, meta-analyses, or subgroup analyses. *JAMA* 2009;302:2361–2362.
43. Balkrishnan R, Furberg CD. Developing an optimal approach to global drug safety. *J Intern Med* 2001;250:271–279.

Chapter 13
Assessment of Health-Related Quality of Life

Michelle J. Naughton, Ph.D.
Sally A. Shumaker, Ph.D.
Guest Contributors

The term "quality of life" is widely used by psychologists, sociologists, economists, policy makers, and others. However, what is meant by quality of life varies greatly depending on the context. In some settings, it may include such components as employment status, income, housing, material possessions, environment, working conditions, or the availability of public services. The kinds of indices that reflect quality of life from a medical or health viewpoint are very different, and would include those aspects that might be influenced not only by conditions or diseases but also by medical treatment or other types of interventions. Thus, the term "health-related quality of life (HRQL)" is now commonly used to mean the measurement of one's life quality from a health or medical perspective.

Components incorporated under the broad rubric of HRQL have been part of clinical trials before this term was established. Measures of physical functioning and psychological functioning, such as depression and anxiety, and a variety of symptoms such as pain, are well-established outcome variables. Negative effects, typically adverse symptoms from various organ systems, are also routinely assessed (see Chap. 12). In addition, some components have for years been among the baseline factors often collected to characterize the study population. However, the introduction of the HRQL concept has provided important additions and refinements in the way clinical trials are designed. A major advance is that more attention is given to the participants and their experiences and perceptions and how these might be affected by the study intervention. The quantification of these measures has become more sophisticated with assistance from investigators trained in measurement theory and psychometrics.

Over the past 20 years, there has been a rapidly growing interest in the inclusion of HRQL measures to assess intervention effects in trials of a variety of interventions [1–6]. In response to this interest, methods to assess health status and HRQL have proliferated, and there are a number of valid, reliable, and sensitive instruments that have been developed for use in clinical trials [2, 6–9]. In this chapter, we provide a definition of HRQL, examine the uses of HRQL assessment in clinical trials, discuss study design considerations, and consider the interpretation of data resulting from these research investigations. In the last section, we provide a brief

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_13, © Springer Science+Business Media, LLC 2010

introduction to utility measures and preference scaling, which although distinct from most psychometrically based HRQL measures, are useful in some types of clinical intervention studies.

Fundamental Point

Assessments of the effects of interventions on participants' health-related quality of life is a critical component of many clinical trials, especially ones which involve interventions directed to the primary or secondary prevention of chronic diseases.

Defining Health-Related Quality of Life

Historically, experts in the field of quality of life have held varying viewpoints on how to define the concept. Recent years, however, have brought greater convergence of opinion with respect to definitions of HRQL [10]. Several definitions have been proposed and these have ranged from very broad perspectives, reminiscent of the early definitions of quality of life, to narrower definitions that are more specific to HRQL [11]. We have adopted a definition of HRQL proposed by Wenger and Furberg [12].

HRQL encompasses: "Those attributes valued by patients, including: their resultant comfort or sense of well-being; the extent to which they were able to maintain reasonable physical, emotional, and intellectual function; and the degree to which they retain their ability to participate in valued activities within the family, in the workplace, and in the community." Explicit within this definition is the multidimensional aspect of HRQL, and that actual functional status and the individuals' perceptions regarding "valued activities" are critical to identify. Although there has been some debate among experts on the definition of HRQL, there is general agreement on the primary dimensions of HRQL that are essential to any HRQL assessment [10]. These fundamental or primary dimensions include: physical functioning, psychological functioning, social functioning and role activities, and the individuals' overall assessment of their life quality and perceptions of their health status (Table 13.1). Thus, most experts agree that in order for investigators to assert that they have measured HRQL in a particular clinical trial, a minimum of these dimensions should be included. However, there are certainly instances in which fewer dimensions may be applicable to a specific intervention or population. For example, it is unlikely that in the examination of the short-term effects of hormone therapy on peri-menopausal symptoms, the general routine physical functioning of the study participants (women in their mid-forties to early fifties) will be influenced. Thus, the inclusion of this dimension of HRQL in the trial may simply increase participant burden without benefit. In such instances, it is important for

Table 13.1 Dimensions of health-related quality of life

Primary dimensions
Physical functioning
Social functioning
Psychological functioning
Perception of overall quality of life
Perceptions of health status
Additional dimensions
Neuropsychological functioning
Personal productivity
Intimacy and sexual functioning
Sleep disturbance
Pain
Symptoms

investigators to indicate clearly the dimensions of HRQL used in the trial and justify their selection of a subset of HRQL dimensions in order to avoid the perception of bias, for example, deleting HRQL dimensions that might make the treatment under study "look bad."

For specific interventions, other commonly assessed dimensions of HRQL may be important (Table 13.1). These include cognitive or neuropsychological functioning, personal productivity, and intimacy and sexual functioning. Measures of sleep disturbance, pain, and symptoms, which are associated with a condition/illness and the adverse effects of treatment, are also often assessed.

Primary HRQL Dimensions

Physical functioning refers to an individual's ability to perform daily life activities. These types of activities are often classified as either "activities of daily living," which include basic self-care activities, such as bathing and dressing, or "intermediate activities of daily living," which refer to a higher level of usual activities, such as cooking, performing household tasks, and ambulation.

Social functioning is defined as a person's ability to interact with family, friends, and the community. Instruments measuring social functioning may include such components as the person's participation in activities with family, friends, and in the community, and the number of individuals in his or her social network. A key aspect of social functioning is the person's ability to maintain social roles and obligations at desired levels. An illness or intervention may be perceived by people as having less of a negative impact on their daily lives if they are able to maintain role functions that are important to them, such as caring for children or grandchildren or engaging in social activities with friends. In contrast, anything that reduces one's ability to participate in desired social activities, even though it may improve clinical status, may reduce the person's general sense of social functioning.

Psychological functioning of a person refers to the individual's emotional well-being. It has been common to assess the negative effects of an illness or intervention, such as levels of anxiety, depression, guilt, and worry. However, the positive emotional states of individuals should not be neglected. Interventions may produce improvements in a person's emotional functioning, and, therefore, such aspects as joy, vigor, and hopefulness for the future are also important to assess.

Overall quality of life represents a person's perception of his or her overall quality of life. For example, participants may be asked to indicate a number between 0 (worst possible quality of life) and 10 (best possible quality of life), which indicates their overall quality for a defined time period (for example, in the last month).

Perceptions of health status need to be distinguished from actual health. Individuals who are ill and perceive themselves as such, may, after a period of adjustment, reset their expectations and adapt to their life situation, resulting in a positive sense of well-being. In contrast, persons in good health may be dissatisfied with their life situation and rate their overall quality of life as poor. Participants may be asked to rate their overall health in the past month, their health compared to others their own age, or their health now compared to 1 year ago. It is interesting to note that perceived health ratings are strongly and independently associated with an increased risk of mortality [13, 14], indicating that health perceptions may be important predictors of health outcomes, independent of clinical health status.

Additional HRQL Dimensions

Neuropsychological functioning refers to the cognitive abilities of a person, such as memory, recognition, spatial, and psychomotor skills. This dimension is being more commonly assessed for a wide range of health conditions or procedures, such as stroke or postcardiac surgery, as well as for studies of older cohorts.

Personal productivity is a term used to encompass the range of both paid and unpaid activities in which individuals engage. Measures of this dimension might include paid employment (for instance, date of return to work, hours worked per week), household tasks, and volunteer or community activities.

Intimacy and sexual functioning refer to one's ability to form and maintain close personal relationships and engage in sexual activities, as desired. Instruments measuring sexual functioning include items regarding a person's ability to perform and/or participate in sexual activities, the types of sexual activities in which one engages, the frequency with which such activities occur, and persons' satisfaction with their sexual functioning or level of activity. This dimension of HRQL is particularly important in studies in which the disease's or condition's natural history, or its treatment, can influence sexual functioning (for example, antihypertensive therapy, prostate cancer surgery, and other forms of cancer therapy).

Sleep disturbance has been related to depression and anxiety, as well as diminished levels of energy and vitality. Instruments assessing sleep habits may examine such factors as sleep patterns (e.g., ability to fall asleep at night, number of times awakened during the night, waking up too early in the morning, or difficulty in waking up in the morning, number of hours slept during a typical night), and the restorativeness of sleep.

Pain is another commonly assessed dimension of HRQL, particularly in such chronic conditions as arthritis or orthopedic injuries. Assessments of pain may include measures of the degree of pain related to specific physical activities, such as bending, reaching, or walking upstairs, as well as the type of pain, such as throbbing, shooting, and aching. The frequency and duration of pain are also generally recorded.

Symptoms associated with study conditions or interventions are an integral part of most clinical trials but are also one aspect of HRQL. By incorporating symptoms into HRQL assessments more systematically, we now have more sophisticated accounts of the frequency of symptoms, the severity of symptoms, and the degree to which symptoms interfere with daily functioning. Symptom checklists are often tailored to the specific condition or illness being studied, and require investigators to have knowledge of common symptoms associated with an illness, the symptoms which may be produced or relieved by an intervention, and the time course in which these symptoms may be expected to occur during the course of the clinical trial. For many conditions, however, there are symptom checklists that have already been validated that investigators should explore prior to developing new symptom measures.

Although all of the above dimensions of HRQL are the most commonly assessed aspects of HRQL, the specific dimensions relevant for a given clinical trial will depend upon the intervention under investigation, the disease or condition being studied, and the study population, which may vary by the age of the participants, their ethnic identity, or cultural background [15].

Uses of Health-Related Quality of Life

For many individuals, there are really only two outcomes that are important when assessing the efficacy of a particular treatment: changes in their life expectancies, and the quality of their remaining years. HRQL provides a method of measuring intervention effects, as well as the effects of the untreated course of diseases, in a manner that makes sense to both the individual and the investigator. As countries where chronic rather than acute conditions dominate the health care system, the major goals of intervention, include: the prevention of disease onset, and when the disease has developed, the reduction of symptoms, maintenance or improvement in functional status, and the potential to prolong life. In interventions for disease prevention, it is reasonable for individuals to expect that interventions, while decreasing the probability that they will develop a chronic condition, will not in the process

significantly reduce their current functioning. In terms of chronic conditions where the goal is generally not a cure, it is important to determine how the person's life is influenced by both the disease and its treatment, and whether the effects of treatment are better or worse than the effects of the course of the underlying disease.

There are now many published studies assessing the quality of life of participants in clinical trials. Some use quality of life as baseline covariates and/or outcome measures of the effect of a trial on a person's life quality. Others may also use baseline quality of life measures to predict patients' overall survival, adherence, and adjustment to treatment and/or the disease itself. Several examples illustrate the value of including quality of life measures in clinical trials. In one early clinical trial, Sugarbaker and colleagues [16] examined 26 participants with soft tissue sarcoma and compared the effects of two treatments on quality of life. Participants were randomized to either amputation plus chemotherapy or limb-sparing surgery plus radiation therapy plus chemotherapy. After treatments had been completed and the participants' physical status had stabilized, economic impact, mobility, pain, sexual relationships, and treatment trauma were assessed. Contrary to expectations, participants receiving amputation plus chemotherapy reported better mobility and sexual functioning than those receiving limb-sparing surgery plus radiation and chemotherapy.

A more recent example from the heart disease literature is the quality of life results from the Women's Health Initiative (WHI) hormone therapy trials. During the 1980s and early 1990s, observational and case-control studies suggested that the use of estrogen would decrease the incidence of cardiovascular events among postmenopausal women. In order to determine if this observation would be replicated in a large, randomized controlled trial, the WHI was initiated in 1994 [17]. Consenting postmenopausal participants ages 50–79 were randomized to either conjugated equine estrogens plus medroxyprogesterone acetate (CEE + MPA) vs. placebo if they had a uterus, or conjugated equine estrogens (CEE-alone) vs. placebo among participants who had been hysterectomized. HRQL was assessed annually after trial initiation. In 2002, the trial testing CEE + MPA was stopped early, due to higher rates of cardiovascular events and breast cancer among women in the CEE + MPA arm vs. the placebo group [18]. A year and a half later, the CEE-alone trial was also stopped due to higher rates of stroke and thromboembolic events among women randomized to the hormone therapy group [19]. The results of these two trials had a major impact on the care recommendations of postmenopausal women, and spurred a debate among primary care practitioners, cardiologists, and gynecologists about the validity of the WHI results [20]. One argument was that although estrogen therapy may not be indicated for cardiovascular disease protection, women still report better quality of life when taking estrogen therapy. The quality of life results from the WHI, however, did not support this argument [21]. Among women randomized to CEE + MPA vs. placebo, active treatment was associated with a statistically significant, but small and not clinically meaningful benefit, in terms of sleep disturbance, physical functioning, and bodily pain 1 year after the initiation of the study. At 3 years, however, there were no significant benefits in terms of any quality of life outcomes. Among women aged 50–54 with

moderate to severe vasomotor symptoms at baseline, active therapy improved vasomotor symptoms and sleep quality, but had no benefit on other quality of life outcomes. Similar results were found in the CEE-alone trial of the WHI among women with hysterectomy. At both 1 year and 3 years after the initiation of the trial, CEE had no significant effect on HRQL [22]. Thus, the potential harmful effects of estrogen therapy among postmenopausal women were not outweighed by gains in quality of life.

Other studies have illustrated the useful inclusion of quality of life measures as secondary outcomes in clinical trials. Kornblith et al. [23] compared azacytidine vs. supportive care in people with myelodysplastic syndrome (MDS). The study showed that participants receiving azacytidine had lower rates of transformation to acute myelogenous leukemia or death than those randomized to supportive care. The quality of life results demonstrated improvements in fatigue, dyspnea, physical functioning, positive affect, and psychological distress among those patients receiving azacytidine. In contrast, participants receiving only supportive care reported stable or declining quality of life over the course of the study. Prior to the trial, allogeneic bone marrow transplantation was the only effective treatment for MDS, but was an option for only a small proportion of patients. The results of the trial led to the establishment of azacytidine as a treatment option for MDS, and illustrated the added benefit of using quality of life and symptom assessments as outcomes.

Methodological Issues

The rationale for a well-designed and conducted randomized clinical trial to assess HRQL measures is the same as for other response variables. Because the data are primarily subjective, special precautions are necessary. A control group allows the investigator to determine which changes can be reasonably attributed to the study intervention. The double-blind design minimizes the effect of investigator bias. The findings will be all the more credible if hypotheses are established a priori in the trial protocol.

The basic principles of data collection (Chap. 11) which ensure that the data are of the highest quality are also applicable. The methods for assessment must be clearly defined. Training sessions of investigators and staff are advisable. Pretesting of forms and questionnaires may enhance user and patient acceptability, and ensure higher quality data. An ongoing monitoring or surveillance system enables prompt corrective action when errors and other problems are found.

Trial Design

Several protocol issues must be taken into account when using HRQL measures in clinical trials, including the time course of the trial, the frequency of contact with

the study participants, the timing of clinical assessments, the complexity of the trial design, the number of participants enrolled, and participant and staff burden. The goal of the HRQL investigation is to incorporate the HRQL measures to the trial protocol without compromising other aspects of the trial design. For example, in the case of a trial design with frequent participant contacts and multiple clinical measures, it may be necessary to focus the assessment of HRQL on a subset of critical dimensions in order to minimize participant and staff burden.

At the same time, however, if a decision to measure HRQL is made, then like other measures, it should be viewed as an important variable in the overall trial design. Reducing its measurement to very brief and potentially less reliable measures, or to only one or two dimensions, may seriously diminish the integrity of the overall study design and yield useless information. For some trials, HRQL will be the primary endpoint, and the focus with respect to staff and patient time should be on the HRQL battery. For instance, if comparing two antihypertensive drugs which have comparable efficacy with respect to blood pressure reduction, but different effects on HRQL, then HRQL should be the critical outcome variable, and the study measures, as well as staff and patient time should reflect this fact [24].

Study Population

It is critical to specify key population demographics that could influence the choice of instruments, the relevant dimensions of HRQL to be assessed, or the mode of administration. Thus, educational level, gender, age range, the language(s) spoken, and cultural diversity should be carefully considered prior to selecting the HRQL battery of measures. It could be that a cohort of patients over the age of 70 may have more vision problems than middle-aged persons, making self-administered questionnaires potentially inadvisable. Ethnically diverse groups also require measures that have been validated across different cultures and/or languages [25].

It is also important to be sensitive to how the disease will progress and affect the HRQL of patients in the control group, as it is to understand the effects of the study intervention. For example, in patients with congestive heart failure assigned to the placebo-control arm of the study, we can expect a worsening of symptoms such as shortness of breath and fatigue, both of which will influence daily functioning. The point is to select dimensions and measures of HRQL that are sufficiently sensitive to detect changes in *both* the treated and the control group patients. Uses of the same instruments for both groups will ensure an unbiased and comparable assessment.

Intervention

Three major intervention-related factors are relevant to HRQL: the positive and adverse effects of treatment, the time course of the effects, and the possible synergism

of the treatment with existing medications and conditions. It is important to understand how a proposed treatment could affect the various dimensions of an individual's life quality in both positive and negative ways. Some oral contraceptives, for instance, may be very effective in preventing pregnancy, while producing aversive symptoms like bloating and breast tenderness.

The time course of an intervention's effects on dimensions of HRQL is also important both in terms of the selection of measures and the timing of when HRQL measures are administered to study participants. In a trial comparing coronary artery bypass graft (CABG) surgery to angioplasty, an assessment of HRQL 1 week postintervention might lead to an interpretation that the surgical arm was more negative than angioplasty for HRQL since the individuals in this arm of the trial would still be recovering from the surgical procedure, and the effects of sore muscles and surgical site discomfort could overwhelm any benefits associated with CABG. However, at 6 months postintervention, the benefits of CABG surgery, such as relief from angina, might be more profound than the benefits received from angioplasty. Thus, the timing of the HRQL assessment may influence how one interprets the benefits (or negative effects) of the interventions.

Furthermore, it is important to know the medications the study population is likely to be on prior to randomization to the study intervention, and how these medications might interact with the trial intervention (either a pharmacological or behavioral intervention) to influence dimensions of HRQL.

Selection of HRQL Instruments

Measures of HRQL can be classified as either generic (that is, instruments designed to assess HRQL in a broad range of populations) or condition/population-specific (instruments designed for specific diseases, conditions, age groups, or ethnic groups) [15]. Within these two categories of measures are single questionnaire items; dimension-specific instruments, which assess a single aspect of HRQL; health profiles, which are single instruments measuring several different dimensions of HRQL; and a battery of measures, which is a group of instruments assessing both single and multiple dimensions of HRQL. In assessing HRQL outcomes, the trend has been toward the use of either profiles or batteries of instruments.

Some of the more commonly used generic HRQL instruments are the SF-36 [26] and the EQ-5D [27]. Frequently used condition-specific instruments include the Functional Assessment of Cancer Therapy (FACT) [28] and the European Organization for Research and Treatment of Cancer Quality of Life (EORTC QLQ) [29], both of which are multidimensional measures assessing the HRQL of individuals with cancer. Other condition specific instruments include the McGill Pain Questionnaire, for the measurement of pain [30]; the Centers for Epidemiological Studies – Depression (CES-D) [31], the Profile of Mood States (POMS) [32], and the Psychological General Well-Being Index (PGWB) [33], all of which assess

psychological distress and well-being; and the Barthel Index to measure physical functioning and independence [34, 35].

The type of instruments selected for inclusion in a clinical trial will depend on the goals of the intervention. Within a given dimension of HRQL, like physical functioning, one can assess the degree to which an individual is able to perform a particular task, his or her satisfaction with the level of performance, the importance to him or her of performing the task, or the frequency with which the task is performed. Thus, the aspects of HRQL measured in clinical trials vary depending on the specific research questions of the trial.

Many investigators have made the mistake of adopting a questionnaire developed for another population only to find that the distribution of responses obtained is skewed. In part, this may be due to volunteers for a trial comprising a select group, often healthier than people in general with the same conditions. This point underscores the need to pretest any proposed instrument before a trial.

There are a range of techniques that have been used to construct HRQL measures. It is beyond the scope of this chapter to review these techniques, but references regarding scaling procedures and psychometric considerations of instruments (reliability, validity, and the responsiveness of instruments to change) may be consulted [3, 6–10, 36, 37]. It is important to note that in selecting HRQL instruments, investigators should be certain of the psychometric integrity of the measures. Fortunately, today there are a number of instruments available that meet the standards put forward by traditional measurement theory.

Modes of Administration

HRQL data can be collected from interviews (telephone or face-to-face), or from self-administered instruments (in-person or by mail). Self-administered instruments are more cost-effective from a staffing perspective, and may yield more disclosure on the part of the participant, particularly with the collection of sensitive information. However, self-administered instruments tend to yield more missing and incomplete data and do not allow for clarification. In the long run, and with some populations, self-administered instruments may actually prove to be more expensive than interviewer administered instruments, if more staff time is needed to follow-up with participants to clarify responses to particular items and/or to attempt to get participants to respond to questionnaire items that were left blank on the survey.

Interviewer administered instruments usually provide more complete data sets and allow for probes and clarification. However, there may be a reluctance on the part of some participants to openly discuss some HRQL issues (for example, depression, sexuality), whereas they may be willing to respond to questions about these same issues in a self-administered format. For populations with a relatively high proportion of functional illiteracy, in-person interviewer administration may be required. Interviewer administration may also be the best way to obtain information for culturally diverse populations. Finally, interviewer administered instruments are

subject to interviewer bias and require intensive interviewer training, certification, and repeat training, especially within the context of multisite clinical trials, which may be of a long duration. Thus, often they can be considerably more expensive than self-administered instruments and serious thought must be given at the planning phases of a trial regarding the trade-offs between these two strategies.

In practice, clinical trials that include HRQL as outcomes usually incorporate a combination of profiles augmented with either generic or population-specific measures of the dimensions most relevant to the study population and intervention. In addition, most HRQL measures are designed to be either interviewer- or self-administered, and both modes of administration can be used within single trials.

Frequency of Assessment (Acute Vs. Chronic)

The frequency with which HRQL will need to be assessed in an intervention will depend on the nature of the condition being investigated (acute vs. chronic) and the expected effects (both positive and negative) of treatment. At a minimum, as with all measurements collected in a clinical trial, a baseline assessment should be completed prior to randomization and the initiation of the intervention. Follow-up HRQL assessments should be timed to match expected changes in functioning due to either the intervention or the condition itself, and study objectives.

In general, acute conditions resolve themselves in one of four ways: a rapid resolution without a return of the condition or symptoms; a rapid resolution with a subsequent return of the conditions after some period of relief (relapse); conversion of the condition to a chronic problem; or death [38]. In the case of rapid resolution, HRQL assessments would likely focus on the participant's symptoms in the short-term, and allow for comparisons between the side effects of treatment that might assist resolution vs. the relative impact of symptoms on the participant's daily life. With respect to an acute condition where there is a risk of relapse (for example, gastric ulcer), a longer duration of follow-up is necessary because relapses can occur frequently and may have a broad impact on the participant's general functioning and well-being.

If the acute problem converts to a chronic condition, the evaluation of adverse symptoms vs. treatment side-effects remains important, but is complicated by the duration of time and the problem of how to balance health outcomes in making treatment decisions. A cancer patient experiencing acute pain, for instance, will often be treated with opiods, where appropriate, despite their negative side-effects. Most patients will gladly accept the negative effects of the drugs (for instance, sedating effect) in exchange for immediate relief from pain. However, if treatment extends for a long period of time, the cumulative effects of sedation and other side effects must be weighed against the benefit of pain control. Interest in HRQL has been greater in the management of chronic conditions, where there is a growing relative emphasis on morbidity over mortality. In chronic diseases, postponement of onset and treatment of associated symptoms may be the most important factors to assess.

Symptom Expression (Episodic Vs. Constant)

Chronic conditions with episodic symptomatic flare-ups (e.g., myasthenia gravis) can mimic acute conditions. However, a major distinction between the two is that often some interventions for the chronic conditions must be administered during the asymptomatic periods. In addition, relief from symptoms from many chronic conditions is not as complete as that for acute conditions which, by definition, resolve in a short time period. If the treatment carries side-effects or adds to unrelated health risks, HRQL assessments ought to be completed during both latent and symptomatic periods in order to better characterize the impact of the condition and intervention on the participants.

Functional Impact (Present Vs. Absent)

For specific conditions which have little or no adverse effect upon patient function, treatments are best evaluated on the basis of their impact on survival. In these situations, HRQL assessments will be of secondary importance. However, when a disease or condition affects functional capacity, treatments for that condition ought to be evaluated for their influence, both positive and negative, upon the participants' level of functioning and well-being. Again, in these situations, the type of HRQL instruments used and the timing of the assessments will depend on the nature of the condition, the treatment, and the expected time course of effects on the participants.

Interpretation

The dimensions composing HRQL are influenced by a broad range of factors. It is important to maintain a distinction between these moderating factors and HRQL. Moderating factors can be divided into three categories: contextual, interpersonal, and intrapersonal [39]. Contextual factors include such variables as the setting (for example, urban–rural, single dwelling building vs. high rise); the economic structure; and sociocultural variations. Interpersonal factors include such variables as the social support available to individuals, stress, economic pressures, and the occurrence of major life events, such as bereavement and the loss of a job. Intrapersonal factors have to do with factors associated with the individual, such as coping skills, personality traits, or physical health. This distinction between the dimensions that comprise HRQL and the factors which moderate HRQL has implications for the selection of HRQL measures in specific trials, as well as data analysis and interpretation.

In addition to these three categories of moderating factors, it is important to realize that any intervention may induce changes, improvements as well as impairments,

in a participant's well-being. Changes in the natural course of the disease or conditions must be considered, especially in trials of relatively long duration. Concomitant interventions or the regimen of care itself may also affect HRQL. This is particularly likely to happen in trials where the active intervention is considerably different than that for the control group. It is important to consider what effects the intervention will have on the participants' well-being before initiating the trial in order to be able to assess the impact of these factors on the HRQL of the participants.

Scoring of HRQL Measures

In most clinical trials, HRQL is assessed by several instruments measuring dimensions of HRQL considered to be critical to the intervention. Scores resulting from these measures are usually calculated by dimensions of HRQL so that a separate score is calculated for physical functioning or social well-being, and so on. Some instruments may also produce an overall HRQL score in addition to separate scores for each dimension (for example, FACT) [28].

Scores resulting from HRQL instruments are used to address specific research questions, most notably, to assess changes in specific HRQL dimensions, throughout the course of the trial; to describe the treatment and control groups at distinct points in time; and to examine the correspondence between HRQL measures and clinical or physiological measures. Plans for data analysis are tailored to the specific goals and research questions of the clinical trial, and a variety of standard statistical techniques are used to analyze HRQL data.

Determining the Clinical Significance of HRQL Measures

An important issue in evaluating HRQL measures is determining how to interpret score changes on a given scale. For example, how many points must one increase or decrease on a scale for that change to be considered clinically meaningful? Does the change in score reflect a small, moderate, or large improvement or deterioration in a participant's health status? Recent years have seen an increase in research examining the question of the clinical significance of HRQL scores. Demonstrating the clinical significance of HRQL measures is also important for achieving successful product claims through regulatory agencies [3].

In order to be acceptable as outcome measures in clinical trials, HRQL instruments must have acceptable reliability, validity, and responsiveness to changes either in clinical status and/or effects of the intervention. Information on how to interpret changes in HRQL scores over time is based on the minimal important difference (MID) [40, 41]. When the change score is connected to clinical anchors, the MID is sometimes referred to as the minimal clinically important difference (MCID). Responsiveness corresponds to the instrument's ability to measure changes, whereas

the MID is defined as the smallest score or change in scores that is perceived by patients as improving or decreasing their HRQL and which would lead a clinician to consider a change in treatment or patient follow-up [41–43]. The responsiveness of a HRQL instrument and the MID can vary based on population and contextual characteristics. Thus, there will not be a single MID value for a HRQL instrument across all uses and patient samples. It is likely that there is a range in MID estimates that vary across patient populations and observational and clinical trial applications [40].

A variety of methods have been used to determine the MID in HRQL instruments. However, there is currently no consensus on which method is best for determining the MID. Therefore, the MID determination should be based on multiple approaches [40, 43]. More in-depth discussion of issues regarding the MID and HRQL and other patient-reported outcome measures can be found elsewhere [40].

Utility Measures/Preference Scaling

The types of HRQL instruments discussed in this chapter have been limited to measures which were derived using psychometric methods. These methods examine the reliability, validity, and responsiveness of instruments. Other approaches to measuring quality of life and health states are used, however, and include utility measures and preference scaling [44]. Utility measures are derived from economic and decision theory, and incorporate the preferences of individuals for particular treatment interventions and health outcomes. Utility scores reflect persons' preferences and values for specific health states and allow morbidity and mortality changes to be combined into a single weighted measure, called "quality-adjusted life years (QALY)" [45, 46]. These measures provide a single summary score representing the net change in the participant's quality of life (the gains from the treatment effect minus the burdens of the side effects of treatment). These scores can be used in cost-effectiveness analyses that combine quality of life and duration of life. Ratios of cost per QALY can be used to decide among competing interventions.

In utility approaches, one or more scaling methods are used to assign a numerical value from 0.0 (death) to 1.0 (full health) to indicate an individual's quality of life. Procedures commonly used to generate utilities are lottery or standard gamble, most usually the risk of death one would be willing to take to improve a state of healthy [46]. Preferences for health states are generated from the general population, clinicians, or patients using multiattribute scales, visual analogue rating scales, time trade-off (how many months or years of life one would be willing to give up in exchange for a better health state), or other scaling methods [44, 47, 48]. Utility measures are useful in decision-making regarding competing treatments and/or for the allocation of limited resources. They also can be used as predictors of health events. Clarke and colleagues examined the use of index scores based on the EQ-5D, a five-item generic health status measure, as an independent predictor of vascular events, other major complications and mortality in people with type 2 diabetes [49]. A cohort of 7,348 participants, aged 50–75, were recruited to the

Fenofibrate Intervention and Event Lowering in Diabetes (FIELD) study. After adjusting for standard risk factors, a 0.1 higher index score derived from the EQ-5D was associated with an additional 7% lower risk of vascular events, a 13% lower risk of diabetic complications, and a 14% lower rate of all-cause mortality.

In general, psychometric and utility-based methods measure different components of health. The two approaches result in different yet related, and complementary assessments of health outcomes, and both are useful in clinical research. Issues regarding the use of utility methods include the methodologies used to derive the valuation of health states; the cognitive complexity of the measurement task, potential population, and contextual effects on utility values; and the analysis and interpretation of utility data [39, 41]. For a further review of issues related to utility analyses/preference scaling, and the relationship between psychometric and utility-based approaches to the measurement of life quality, additional references may be consulted [44–50].

References

1. Quality of Life Assessment in Cancer Clinical Trials. Report of the Workshop on Quality of Life Research in Cancer Clinical Trials. *USDHHS*, Bethesda, Maryland, 1991.
2. Spilker B (ed.). *Quality of Life and Pharmacoeconomics in Clinical Trials.* Philadelphia: Lippincott-Raven Publishers, 1996.
3. Revicki DA, Osoba D, Fairclough D, et al. Recommendations on health-related quality of life research to support labeling and promotional claims in the United States. *Qual Life Res* 2000;9:887–900.
4. Fairclough DL. *Design and Analysis of Quality of Life Studies in Clinical Trials (Interdisciplinary Statistics).* Boca Raton, Florida: Chapman & Hall/CRC, 2002.
5. Fayers P, Machin D. *Quality of Life: The Assessment, Analysis and Interpretation of Patient-Reported Outcomes*. Chichester: John Wiley & Sons, Ltd, 2007.
6. Ganz PA, Reeve BB, Clauser SB, Lipscomb J. Patient-reported outcomes assessment in cancer trials. *J Clin Oncol* 2007;25:5049–5141.
7. Stewart AL, Ware JE (eds.). *Measuring Functioning and Well-Being.* Durham, NC: Duke University Press, 1992.
8. Wilkin D, Hallam L, Doggett M. *Measures of Need and Outcome for Primary Health Care.* New York: Oxford Medical Publications, 1992.
9. McDowell I. *Measuring Health: A Guide to Rating Scales and Questionnaires.* New York: Oxford University Press, 2006.
10. Berzon R, Hays RD, Shumaker SA. International use, application and performance of health-related quality of life instruments. *Qual Life Res* 1993;2:367–368.
11. Stewart A. Conceptual and methodologic issues in defining quality of life: State of the art. *Prog Cardiovasc Nurs* 1992;7:3–11.
12. Wenger NK, Furberg CD. Cardiovascular Disorders. In Spilker B (ed.). *Quality of Life Assessment in Clinical Trials.* New York: Raven Press, 1990.
13. Mossey JM, Shapiro E. Self-rated health: A predictor of mortality among the elderly. *Am J Public Health* 1982;72:800–808.
14. Kaplan GA, Camacho T. Perceived health and mortality: A nine-year follow-up of the human population laboratory cohort. *Am J Epidemiol* 1993;117:292–304.
15. Schron EB, Shumaker SA. The integration of health quality of life in clinical research: Experiences from cardiovascular clinical trials. *Prog Cardiovasc Nurs* 1992;7:21–28.

16. Sugarbaker PH, Barofsky I, Rosenberg SA, Gianola FJ. Quality of life assessment of patients in extremity sarcoma clinical trials. *Surgery* 1982;91:17–23.
17. The Women's Health Initiative Study Group. Design of the Women's Health Initiative clinical trial and observational study. *Control Clin Trials* 1998;19:61–109.
18. Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women. *JAMA* 2002;288:321–333.
19. The Women's Health Initiative Steering Committee. Effects of conjugated equine estrogen in postmenopausal women with hysterectomy. *JAMA* 2004;291:1701–1712.
20. Naughton MJ, Jones AS, Shumaker SA. When practices, promises, profits, and policies outpace hard evidence: The post-menopausal hormone debate. *J Soc Issues* 2005;61:159–179.
21. Hays J, Ockene JK, Brunner RL, et al. for the Women's Health Initiative Investigators. Effects of estrogen plus progestin on health-related quality of life. *N Engl J Med* 2003;348:1839–1854.
22. Brunner RL, Gass M, Aragaki A, et al. for the Women's Health Initiative Investigators. Effects of conjugated equine estrogen on health-related quality of life in postmenopausal women with hysterectomy: Results from the Women's Health Initiative Randomized Clinical Trial. *Arch Intern Med* 2005;165:1976–1986.
23. Kornblith AB, Herndon JE, Silverman LR, et al. Impact of azacytidine on the quality of life of patients with myelodysplastic syndrome treated in a randomized phase III trial: A Cancer and Leukemia Group B study. *J Clin Oncol* 2002;29:2441–2452.
24. Testa MA, Anderson RB, Nackley JF, et al. Quality of life and antihypertensive therapy in men: A comparison of captopril and enalapril. *N Engl J Med* 1993;328:901–913.
25. Shumaker SA, Anderson R, Berzon R, Hayes R (eds.). Special Issue. International use, application and performance of health-related quality of life measures. *Qual Life Res* 1993;2:376–495.
26. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). 1. Conceptual framework and item selection. *Med Care* 1992;30:473–483.
27. Rabin R, de Charro F. EQ-5D: A measure of health status from the EuroQol Group. *Ann Med* 2001;33:337–343.
28. Cella DF, Tulsky DS, Gray G, et al. The functional assessment of cancer therapy scale: Development and validation of the general measure. *J Clin Oncol* 1993;11:570–579.
29. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst* 1993;85:365–376.
30. Melzack R. The McGill Pain Questionnaire: Major properties and scoring methods. *Pain* 1975;1:277–299.
31. Radloff LS. The CES-D Scale: A self-report depression scale for research in the general population. *Appl Psych Meas* 1977;1:385–401.
32. McNair DM, Loor M, Droppleman LF. *Manual of the Profile of Mood States.* San Diego, California: Educational and Industrial Testing Service, 1971.
33. Dupuy HJ. The Psychological General Well-Being (PGWB) Index. In Wenger NK, Mattson ME, Furberg CD, Elinson J (eds.). *Assessment of Quality of Life in Clinical Trials in Cardiovascular Therapies.* Washington, DC: Le Jacq Publishing, 1984, pp. 170–183.
34. Granger CV, Albrecht GL, Hamilton BB. Outcomes of comprehensive medical rehabilitation: Mesurement by PULSES profile and the Barthel Index. *Arch Phys Med Rehab* 1979;60: 145–154.
35. Collin C, Wade DT, Davies S, Horne V. The Barthel ADL Index: A reliability study. *Intern Disabil Stud* 1988;10:61–63.
36. Hays RD, Hadorn D. Responsiveness to change: An aspect of validity, not a separate dimension. *Qual Life Res* 1992;1:73–75.
37. Hays RD, Revicki DA. Reliability and Validity (Including Responsiveness). In Fayers P, Hays R (eds.). *Assessing Quality of Life in Clinical Trials* (2nd edition). New York: Oxford University Press, 2005.

38. Cella DF, Wiklund I, Shumaker SA, et al. Integrating health-related quality of life into cross-national clinical trials. *Qual Life Res* 1993;2:433–440.
39. Naughton MJ, Shumaker SA, Anderson R, Czajkowski S. Psychological Aspects of Health-Related Quality of Life Measurement: Tests and Scales. In Spilker B (ed.). *Quality of Life and Pharmacoeconomics in Clinical Trials*. Philadelphia: Lippincott-Raven Publishers, 1996.
40. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008;61:102–109.
41. Jaeschke R, Singer J, Guyatt G. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials* 1991;12:266S–269S.
42. Guyatt G, Walter S, Norman G. Measuring change over time: Assessment the usefulness of evaluative instruments. *J Chronic Dis* 1987;40:171–178.
43. Guyatt G, Osoba D, Wu AW. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002;77:371–383.
44. Weinstein MC, Torrance G, McGuire A. QALYs: The basics. *Value Health* 2009;12:S5–S9.
45. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med* 1993;118:622–629.
46. Revicki DA, Kaplan RM. Relationship between psychometric and utility-based approaches to the measurement of health-related quality of life. *Qual Life Res* 1993;2:477–487.
47. Torrance GW. Integrating Economic Evaluations and Quality of Life Assessments. In Spilker B (ed.). *Quality of Life and Pharmacoeconomics in Clinical Trials*. Philadelphia: Lippincott-Raven Press, 1996.
48. Kaplan RM, Feeny D, Revicki DA. Methods for assessing relative importance in preference based outcome measures. *Qual Life Res* 1993;2:467–475.
49. Clark PM, Hayes AJ, Glasziou PG, et al. Using the EQ-5D index score as a predictor of outcomes in patients with type 2 diabetes. *Med Care* 2009;47:61–68.
50. Special Issue. Moving the QALY forward: Building a pragmatic road. *Value Health* 2009;12:S1–S39.

Chapter 14
Participant Adherence

The terms compliance and adherence are often used interchangeably. Compliance is defined as "the extent to which a person's behavior (in terms of taking medications, following diets, or executing lifestyle changes) coincides with medical or health advice" [1]. The term adherence is defined similarly, but implies active participant involvement. This book uses the term adherence. For example, an adherer is a participant who meets the standards of adherence as established by the investigator. In a drug trial, he may be a participant who takes at least a predetermined amount such as 80% of the protocol dose. There should also be a maximum dose that defines adherence. This dose will depend on the nature of the drug being evaluated (no more than 100% for some, perhaps a bit higher for others). A review of 192 clinical trial publications from high-impact journals reveals important variability in the definition (and reporting) of medication adherence [2].

Medication adherence is a major problem in the practice of medicine. As many as one-third of all prescriptions are reportedly never filled and, among those filled, a large proportion is associated with incorrect administration [3]. Nonadherence has been estimated to cause nearly 125,000 deaths per year and has been linked to 10% of hospital admissions and 23% of nursing home admissions [3]. The direct cost to the US health care system is estimated to be $100 billion. Although these problems apply to clinical practice, low adherence is also a major challenge for clinical trial investigators.

The optimal study from an adherence point of view is one in which the investigator has total control over the participant, the administration of the intervention regimen, which may be a drug, diet, exercise, or other intervention, and follow-up. That situation can only realistically be achieved in animal experiments. Any clinical trial, which, according to the definition in this text, must involve human beings, is likely to have in practice less than 100% adherence with the intervention and the study procedures. There are several reasons for low adherence. Life events such as illnesses, loss of employment, and divorce are important factors leading to reduced adherence. In addition, participants may not perceive any treatment benefit, they may be unwilling to change their behaviors, they are forgetful, may lack family support, or ultimately they may change their minds regarding trial participation. Another reason for low adherence is adverse effects to the medication or intervention. Therefore, even studies of a one-time intervention such as surgery or a single medication dose can suffer from

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_14, © Springer Science+Business Media, LLC 2010

nonadherence. In fact, some surgical procedures can be declined or be reversed. In addition, the participant's condition may deteriorate, and thus require termination of the study treatment or a switch from control to intervention. In a clinical trial in stable coronary disease, participants were randomized to percutaneous coronary intervention (PCI) plus optimal medical therapy or optimal medical therapy alone [4]. Among the 1,149 participants in the PCI group, 46 never underwent the procedure and another 27 had lesions that could not be opened. During a median follow-up of 4.6 years, 32.6% of the 1,138 participants in the optimal medical therapy alone group had revascularization. The trial showed no difference for the primary outcome of all-cause mortality or nonfatal myocardial infarction. However, it is difficult to determine how much the cross-overs influenced the overall finding.

Most of the available information on adherence is obtained from the clinical therapeutic situation rather than from the clinical trial setting. Although the differences between patients and volunteer participants are important, tending to minimize low adherence rates in trials, the basic principles apply to both practice and research. Obviously, the results of a trial can be affected by low adherence with the intervention. It leads to an underestimation of possible therapeutic as well as potential toxic effects and can undermine even a properly designed study. Data from a meta-analysis suggest that the difference in health benefits between high and low adherence is 26% [5]. Given the intention-to-treat principle of analysis (see Chap. 17), in order to maintain equivalent power, a 20% reduction in drug adherence may result in the need for a greater than 50% increase in sample size and 30% reduction will require doubling of the study cohort (see Chap. 8).

This chapter discusses what can be done before enrollment to reduce future adherence problems, how to maintain good adherence during a trial, how to monitor adherence, and how to deal with low adherence. In the monitoring section, we also discuss visit adherence. Readers interested in a more detailed discussion of various adherence issues are referred to an excellent text [6] and a recent review of the literature [7].

Fundamental Point

Many potential adherence problems can be prevented or minimized before participant enrollment. Once a participant is enrolled, taking measures to enhance and monitor participant adherence is essential.

Since reduced adherence with the intervention has a major impact on the power of a trial, realistic estimates of cross-overs, drop-ins and drop-outs must be used in calculating the sample size. Underestimates are common and lead to underpowered trials that fail to test properly the trial hypotheses. See Chap. 8 for further discussion of the sample size implications of low adherence.

A *cross-over* is a participant who, although assigned to the control group, follows the intervention regimen; or a participant who, assigned to an intervention group, follows either the control regimen or the regimen of another intervention group when more than one intervention is being evaluated. A *drop-in* is a special kind of

cross-over. In particular, the drop-in is unidirectional, referring to a person who was assigned to the control group but begins following the intervention regimen. A *drop-out* is also unidirectional and refers to a person assigned to an intervention group who fails to adhere to the intervention regimen. If the control group is either on placebo or on no standard intervention or therapy, as is the case in superiority trials, then the drop-out is equivalent to a cross-over. However, if the control group is assigned to an alternative therapy, as is the case in noninferiority trials, then a drop-out from an intervention group does not necessarily begin following the control regimen. Moreover, in this circumstance, there may be a drop-out from the control group. Participants who are unwilling or unable to return for follow-up visits represent another type of low adherence, sometimes also referred to as drop-outs. Because of the possible confusion in meanings, this text will limit the term drop-out to mean the previously defined adherence-related behavior. Those who stop participating in a trial will be referred to as withdrawals.

Considerations Before Participant Enrollment

There are three major considerations affecting adherence to the study medications that investigators and sponsors ought to consider during the planning phase. First, in selecting the study population, steps should be taken to avoid, to the extent possible, enrollment of study participants who are likely to have low adherence. Second, efforts should be made to limit the impact of design features that may adversely influence the level of adherence. Third, the research setting influences participant adherence over the long term. It is important to have realistic estimates of the adherence level during a trial so that proper upward adjustments of the sample size can be made during the planning phase.

Design Factors

The *study duration* also influences adherence. The shorter the trial, the more likely participants are to adhere with the intervention regimen. A study started and completed in 1 day or during a hospital stay has great advantages over longer trials. Trials in which the participants are under supervision, such as hospital-based ones, tend to have fewer problems with low adherence [8]. It is important to be mindful of the fact that there is a difference between special hospital wards and clinics with trained attendants who are familiar with research requirements and general medical or surgical wards and clinics, where research experience might not be common or protocol requirements might not be appreciated. Regular hospital staff have many other duties which compete for their attention, and they perhaps have little understanding of the need for precisely following a study protocol and the importance of good adherence.

Whenever the study involves participants who will be living at home, the chances for low adherence increase. Studies of interventions that require changing a habit are particularly susceptible to this hazard. A challenge is dietary studies.

A participant may need special meals, which are different from those consumed by other family members. It may be difficult to adhere when having meals outside the home. Multiple educational sessions, including the preparation of meals may be necessary. Family involvement is essential, especially if the participant is not preparing the meals [9, 10]. In studies, when the participants' sources of food come only from the hospital kitchen or are supplied by the trial through a special commissary [11], participants are more likely to adhere with the study regimen than when they buy and cook their own food. This may also allow for blinded design.

Simplicity of intervention is an important factor. Single daily dose drug regimens are preferable to multiple daily dose regimens. Despite a simple regimen, 10–40% of participants have imperfect dosing [7]. A review of 76 trials, in which electronic monitors were used, showed that adherence is inversely proportional to the frequency of dosing [12]. Patients on a four-times-a-day regimen achieved on-schedule average adherence rates of about 50%. Adhering to multiple study interventions simultaneously poses special difficulties. For example, quitting smoking, losing weight, and reducing the intake of saturated fat at the same time require highly motivated participants. Unlike on-going interventions such as drugs, diet, or exercise, surgery and vaccination generally have the design advantage, with few exceptions, of enforcing adherence with the intervention.

Where feasible, a *run-in* period before actual randomization may be considered to identify those potential participants who are likely to become low adherers and thereby exclude them from long-term trials. During the run-in, potential participants may be given either active medication or placebo over several weeks or months. An active run-in allows identification of potential participants who do not have a favorable response to treatment on a biomarker or who develop side effects prior to randomization [13]. A placebo run-in allows a determination of the potential participant's commitment to the study. Run-in phases are common. A 2001 literature search resulted in more than 1,100 examples of trials in which run-in phases were used [14]. This approach was successfully employed in a trial of aspirin and beta-carotene in US physicians [15]. By excluding physicians who reported taking less than 50% of the study pills, the investigators were able to randomize excellent adherers. After 5 years of follow-up, over 90% of those allocated to aspirin reported still taking the pills. An additional goal of the run-in is to stabilize the potential trial participants on specific treatment regimens or to wash-out the effects of discontinued medications. Though the number of participants eliminated by the run-in period is usually small (5–10%), it can be important, as even this level of low adherence affects study power. A potential disadvantage of a run-in is that participants may notice a change in their medication following randomization thereby influencing the blindness of assignment. It also delays entry of participants into a trial by a few weeks.

Berger et al. [14] raised the issue of external validity of the findings of trials that excluded potential low adherers during a run-in phase. Can the results from trials with run-in selection of participants reasonably be fully extrapolated to all those patients meeting the trial eligibility criteria? As always, whether to use a run-in depends on the question being posed. Does the trial have many exclusion criteria

(a so-called efficacy trial) or few exclusions (a pragmatic or effectiveness trial)? Or stated differently – What is the effect of the intervention in optimal circumstances? Or, what is the effect when, as is common in clinical settings, a large number of people fail to adhere to prescribed medication? Both are valid questions, but in the latter situation, as noted earlier, a larger sample size will be required. Lee et al. [16] compared the effect size in 43 clinical trials of selective serotonin uptake inhibitors in patients with depression that included a placebo run-in and those that did not and found no statistically significant difference in the results.

In another approach, the investigator may instruct prospective participants to refrain from taking the active agent and then evaluate how well his request was followed. In the Aspirin Myocardial Infarction Study, for instance, urinary salicylates were monitored before enrollment, and very few participants were excluded because of a positive urine test.

Participant Factors

An important factor in preventing adherence problems before enrollment is the *selection of appropriate participants.* Ideally, only those people likely to follow the study protocol should be enrolled. In the ACCORD trial, the screenees' willingness to test blood sugars frequently was taken as a measure of commitment to participate [17]. This may, however, influence the ability to generalize the findings (see Chap. 3). Several articles have reported that there is convincing evidence that nonadherers are substantially different from adherers in ways that are quite independent of the effects of the treatment prescribed [7, 18–20].

It is usually advisable to exclude certain types of people from participation in a trial. Unless the trial is aimed at people with depression or alcohol or drug addiction, individuals with psychological problems, particularly depressive symptoms and those abusing drugs or alcohol are unlikely to be good participants. Those with cognitive impairment or low literacy may also have more problems with adherence. Similarly, those with a known history of missed appointments or adherence problems ought to be excluded. Logistic factors may also influence adherence, for example, persons who live too far away, or those who are likely to move before the scheduled termination of the trial. Traveling long distances may be an undue burden on disabled people. Those with concomitant disease may be less adherent because they have other medicines to take or are participating in other trials. Furthermore, there is the potential for contamination of the study results by these other medicines or trials. The factors discussed above should, when applied, be incorporated in the study exclusion criteria. They are difficult to define, so the final decision often is left to the discretion of the study investigator.

An *informed participant* appears to be a better adherer. Proper education of the participant is thought to be the most positive factor to high adherence. Therefore, for scientific as well as ethical concerns, the participant (or, in special circumstances,

his guardian) in any trial should be clearly instructed about the study and told what is expected from him. He should have proper insight into his illness and be given a full disclosure of the potential effects – good and bad – of the study medication. Sufficient time should be spent with a candidate and he should be encouraged to consult with his family or private physician. A brochure with information concerning the study is often helpful. As an example, the pamphlet used in the NIH-sponsored Women's Health Initiative trial is shown in Box 14.1. One approach is so-called E-Health Strategies, which refers to information and health services

Box 14.1 Women's Health Initiative Brochure

What is the Women's Health Initiative?

The Women's Health Initiative (WHI) is a major research study of women and their health. It will help decide how diet, hormone therapy, and calcium and vitamin D might prevent heart disease, cancer, and bone fractures. This is the first such study to examine the health of a very large number of women over a long period of time. About 160,000 women of various racial and ethnic backgrounds from 45 communities across the United States will take part in the study.

Who can join the WHI?

You may be able to join if you are

- A woman 50–79 years old
- Past menopause or the "change of life"
- Planning to live in the same area for at least 3 years

Why is this study important?

Few studies have focused on health concerns unique to women. Being a part of this important project will help you learn more about your own health. You will also help doctors develop better ways to treat all women. This study may help us learn how to prevent the major causes of death and poor health in women: heart disease, cancer, and bone fractures.

What will I be asked to do?

If you agree to join us, you will be scheduled for several study visits. These visits will include questions on your medical history and general health habits, a brief physical exam, and some blood tests. Based on your result, you may be able to join at least one of the following programs.

(continued)

Box 14.1 (continued)

- *Dietary.* In this program, you are asked to follow either your usual eating pattern or a low-fat eating program.
- *Hormone.* In this program, you are asked to take either hormone pills or inactive pills (placebos). If you are on hormones now, you would need to talk with your doctor about joining this program.
- *Calcium and Vitamin D.* In this program, you are asked to take either calcium and vitamin D or inactive pills. Only women in the Dietary or Hormone programs may join this program.
- *Health Tracking.* If you are not able to join the other programs, your medical history and health habits will be followed during the study.

How long will the study last?

You will be in the study for a total of 8–12 years, depending on what year you enter the study. This period of time is necessary to study the long-term effects of the programs.

How will I benefit?

If you join the study, your health will be followed by the staff at our center. Certain routine tests will be provided although these are not meant to replace your usual health care. Depending on which program you join, you may receive other health-care services, such as study pills and dietary sessions. You will not have to pay for any study visits, tests, or pills.

You will also have the personal satisfaction of knowing that results from the WHI may help improve your health and the health of women for generations to come.

delivered or enhanced through the internet [21]. Many clinical trials develop websites with educational material directed at physicians and potential participants.

Some investigators have advocated the "talk back" method. It is known that patients, on average, have forgotten half of what they are told when they leave the physician's office. A study of diabetic patients with low health literacy showed that they remembered much better when they were asked to repeat what the physician said [22]. If the investigator says to the study participant that he has high blood pressure that needs treatment, the participant would say, "I have high blood pressure that needs treatment." When told to take one pill every morning until the next clinic visit, the participant would repeat, "I should take one pill every morning until I return for my next clinic visit."

Family support and involvement have emerged as major determinants of adherence. Thus, it is recommended that family members, significant others or friends be informed about the trial and its expectations. After all, a large proportion of participants

Table 14.1 Factors associated with low adherence (adapted from ref. [20])

Unsatisfying participant–investigator relationship
Complexity of drug regimen
Drug effects
Drug–age pharmacokinetics
Adverse effects
Lack of information and inadequate instructions
Personal and cultural beliefs
Health literacy
Functional performance changes
Visual changes
Hearing changes
Cognitive alterations
Emotional health: depression
Lack of social network and support
Limited financial resources

join trials at the support of family and friends [23]. The support they can offer in terms of assistance, encouragement and supervision can be very valuable. This is especially important in trials of lifestyle interventions. For example, cooking classes for spouses as well as participants have been very effective in dietary intervention trials [9, 10].

A large number of factors associated with low adherence have been reported (Table 14.1). Most of them are, as would be expected, the reverse of factors associated with high adherence. There are also factors with no proven association with adherence. The consensus is that age, gender and race or ethnicity have no or a very weak association with adherence.

Maintaining Good Participant Adherence

The foundation for high adherence during a trial is a well-functioning setting with committed clinic staff. Establishing a positive research setting at the first contact with future participants is a worthwhile investment for the simple reason that satisfied participants are better adherers. A warm and friendly relationship between participants and staff established during the recruitment phase should be nurtured. This approach covers the spectrum from trusting interactions, adequate time to discuss complaints, demonstrating sincere concern and empathy, when appropriate, convenient clinic environment, short waiting times, etc. “Bonding” between the participant and clinical trials staff members is a recognized and powerful force in maintaining good adherence. The clinic visits should be pleasant, and participants should be encouraged to contact staff between scheduled visits if they have questions or concern. Close personal contact is key. Clinic staff may make frequent use of the telephone, the mail, and the e-mail. Sending cards on special occasions such as birthdays and holidays is a

helpful gesture. Visiting the participant if he is hospitalized demonstrates concern. It is helpful to investigators and staff to make notes of what participants tell them about their families, hobbies, and work so that in subsequent visits, they can show interest and involvement. Other valued factors are free parking and, for working participants, opportunities for evening or weekend visits. For participants with difficulties attending clinic visits, home visits by staff could be attempted. Continuity of care is ranked as a high priority by participants. Continued family involvement is especially important during the follow-up phase.

During the conduct of a trial, it is important to keep the participants informed about published findings from related trials. They should also be reminded, when applicable, that a monitoring committee is watching the trial data for safety and efficacy. Brief communications from this committee assuring the participants that no safety concern has been noted can also be helpful.

The use of various types of reminders can also reduce the risk of low adherence. Clinic staff should typically *remind* the participant of upcoming clinic visits or study procedures. Sending out postcards, calling, or e-mailing a few days before a scheduled visit can help. Paper-based reminders seem to be most effective [24]. A telephone call though has the obvious advantage that immediate feedback is obtained and a visit can be rescheduled if necessary – a process that reduces the number of participants who fail to keep appointments. Telephoning also helps to identify a participant who is ambivalent regarding his continued participation or who has suffered a study event. To preclude the clinic staff's imposing on a participant, it helps to ask in advance if the participant objects to being called frequently. Asking a participant about the best time to contact him is usually appreciated. Reminders can then be adjusted to his particular situation. In cases where participants are reluctant to come to clinics, more than one staff person might contact the participant. For example, the physician investigator could have more influence with the participant than the staff member who usually schedules visits. In summary, the quantity and quality of interaction between an investigator and the participant can positively influence adherence.

For drug studies, special *pill boxes* help the participant keep track of when to take the medication. These boxes allow participants to divide, by day, all medications prescribed during a 7-day period. If the participant cannot remember whether he took the morning dose, he can easily find out by checking the compartment of the pill box for that day. Special reminders such as noticeable stickers in the bathroom or the refrigerator door or on watches have been used. Placing the pill bottles on the kitchen table or nightstand are other suggestions for participants.

A large number of clinical trials designed to enhance medication adherence have been reported. The interventions have been either behavioral, educational, or both. Several meta-analyses of this type of trial have been published. Based on a review of 61 trials, Peterson [3] concluded that behavioral and educational interventions only led to small improvements in the range of 4–11% in terms of medication adherence. Mail reminders had the largest effect. Another review of 37 trials designed to improve medication adherence in chronic medical conditions reported significant improvement in 20 studies [25]. Most effective were interventions that decreased dosing demands and those involving monitoring and feedback. Few interventions affected the clinical

outcomes. However, a recent Cochrane review [26] concluded that for short-term treatments, simple interventions may increase medication adherence and improve participant outcomes. The methods for improving adherence to chronic conditions are not very effective. The authors call for more fundamental and applied research. A lipid-lowering trial of 13,000 participants showed that repeated telephone and postal reminders only had a nonsignificant improvement in medication adherence [27].

Interventions to maintain good adherence for lifestyle changes can be very challenging [6]. Most people have good intentions that can wane with time unless there is re-enforcement. A special brochure, which contains essential information and reminders, may be helpful in maintaining good participant adherence (Box 14.2). The telephone number where the investigator or staff can be reached should be included in the brochure.

Box 14.2 Aspirin Myocardial Infarction Study Brochure

Text of brochure used to promote participant adherence in the Aspirin Myocardial Infarction Study. DHEW Publication No. (NIH) 76-1080.

1. *Your Participation in the Aspirin Myocardial Infarction Study (AMIS) is Appreciated!* AMIS, a collaborative study supported by the National Heart and Lung Institute, is being undertaken at 30 clinics throughout the United States and involves over 4,000 volunteers. As you know, this study is trying to determine whether aspirin will decrease the risk of recurrent heart attacks. It is hoped that you will personally benefit from your participation in the study and that many other people with coronary heart disease may also greatly benefit from your contribution.
2. *Your Full Cooperation is Very Important to the Study.* We hope that you will follow all clinic recommendations contained in this brochure so that working together, we may obtain the most accurate results. If anything is not clear, please ask your AMIS Clinic Physician or Coordinator to clarify it for you. *Do not hesitate to ask questions.*
3. *Keep Appointments.* The periodic follow-up examinations are very important. If you are not able to keep a scheduled appointment, call the Clinic Coordinator as soon as possible and make a new appointment. It is also important that the dietary instructions you have received be followed carefully on the day the blood samples are drawn. At the annual visit, you must be *fasting.* At the nonannual visits, you are allowed to have a *fat-free diet.* Follow the directions on your Dietary Instruction Sheet. *Don't forget to take your study medication as usual on the day of your visit.*
4. *Change in Residence.* If you are moving within the Clinic area, please let the Clinic Coordinator know of your change of address and telephone number as soon as possible. If you are moving away from the Clinic area,

(continued)

Box 14.2 (continued)

every effort will be made to arrange for continued follow-up here or at another participating AMIS clinic.

5. *Long Vacations.* If you are planning to leave your Clinic area for an extended period of time, let the Clinic Coordinator know so that you can be provided with sufficient study medication. Also give the Clinic Coordinator your address and telephone number so that you can be reached if necessary.
6. *New Drugs.* During your participation in AMIS, you have agreed not to use nonstudy prescribed aspirin or aspirin-containing drugs. Therefore, please call the Clinic Coordinator before starting any new drug as it might interfere with study results. At least 400 drugs contain aspirin, among them cold and cough medicines, pain relievers, ointments and salves, as well as many prescribed drugs. Many of these medications may not be labeled as to whether or not they contain aspirin or aspirin-related components. To be sure, give the Clinic Coordinator a call.
7. *Aspirin-Free Medication.* Your Clinic will give you aspirin-free medication for headaches, other pains, and fever at no cost. The following two types may be provided.
   (a) Acetaminophen. The effects of this drug on headaches, pain, and fever resemble those of aspirin. The recommended dose is 1–2 tablets every 6 h as needed or as recommended by your Clinic Physician.
   (b) Propoxyphene hydrochloride. The drug has an aspirin-like effect on pain only and cannot be used for the control of fever. The recommended dose is 1–2 capsules every 6 h as needed or as recommended by your Clinic Physician.
8. *Study Medication.* You will be receiving study medication from your Clinic. You are to take two capsules each day unless prescribed otherwise. Should you forget to take your morning capsule, take it later during the day. Should you forget the evening dose, you can take it at bedtime with a glass of water or milk. The general rule is: *Do not take more than two capsules a day.*
9. *Under Certain Circumstances It Will Be Necessary to Stop Taking the Study Medication*:
   (a) If you are hospitalized, stop taking the medication for the period of time you are in the hospital. Let the Clinic Coordinator know. After you leave the hospital, a schedule will be established for resuming medication, if it is appropriate to do so.
   (b) If you are scheduled for surgery, we recommend that you stop taking your study medication 7 days prior to the day of the operation. This is because aspirin may, on rare occasions, lead to increased bleeding during surgery. In case you learn of the plans for surgery less than 7 days before it is scheduled, we recommend that you stop the study medication as soon as possible. And again, please let the Clinic

(continued)

Box 14.2 (continued)

Coordinator know. After you leave the hospital, a schedule will be established for resuming medication, if it is appropriate to do so.

(c) If you are prescribed nonstudy aspirin or drugs containing aspirin by your private physician, stop taking the study medication. Study medication will be resumed when these drugs are discontinued. Let the Clinic Coordinator know.

(d) If you are prescribed anticoagulants (blood thinners), discontinue study medication and let your Clinic Coordinator know.

(e) If you have any adverse side effects which you think might be due to the study medication, stop taking it and call the Clinic Coordinator immediately.

10. *Study-Related Problems or Questions.* Should you, your spouse, or anyone in your family have any questions about your participation in AMIS, your Clinic will be happy to answer them. The clinic would like for you or anyone in your family to call if you have any side effects that you suspect are caused by your study medication and also if there is any change in your medical status, for example, should you be hospitalized.
11. *Your Clinic Phone Number Is on the Back of This Brochure. Please Keep This Brochure as a Reference Until the End of the Study.*

Adherence Monitoring

Monitoring adherence is important in a clinical trial for two reasons: first, to identify any problems so that steps can be taken to enhance adherence; second, to be able to relate the trial findings to the level of adherence. In general, analysis of trial outcomes by level of adherence is strongly discouraged (see Chap. 17). However, in so far as the control group is not truly a control and the intervention group is not being treated as intended, group differences are diluted, and generally lead to an underestimate of both the therapeutic effect and the adverse effects. Differential adherence to two equally effective regimens can also lead to possibly erroneous conclusions about the effects of the intervention. The level of adherence that occurred can also be compared with what was expected when the trial was designed.

In some studies, measuring adherence is relatively easy. This is true for trials in which one group receives surgery and the other group does not, or for trials which require only a one-time intervention. Most of the time, however, assessment of adherence is more complex. No single measure of adherence gives a complete picture, and all are subject to possible inaccuracies and varying interpretations. Furthermore, there is no widely accepted definition or criterion for either high or low adherence [28, 29]. A review of 192 publications showed that only 36% assessed and reported

medication adherence [2]. The level of adherence that occurred can also be compared to what was expected when the trial was designed.

In monitoring adherence for a long-term trial, the investigator may also be interested in changes over time. When reductions in adherence are noted, corrective action can possibly be taken. This monitoring could be by calendar time (e.g., current 6 months versus previous 6 months) or by clinic visit (e.g., follow-up visit number 4 vs. previous visits). In multicenter trials, adherence to the intervention also ought to be examined by clinic. In all studies, it is important for clinic staff to receive feedback about level of adherence. In double-blind trials where data by study group generally should not be disclosed, the adherence data can be combined for the study groups. In trials that are not double-blind, all adherence tables can be reviewed with the clinic staff. Frequent determinations obviously have more value than infrequent ones. A better indication of true adherence can be obtained. Moreover, when the participant is aware that he is being monitored, frequent measures may encourage adherence.

There are several indirect methods of assessing adherence. In drug trials, *pill or capsule count,* is the easiest and most commonly used way of evaluating participant adherence. Since this assumes that the participant has ingested all medication not returned to the clinic, the validity of pill count is debated. For example, if the participant returns the appropriate number of leftover pills at a follow-up visit, did he in fact take what he was supposed to, or take only some and throw the rest out? Pill count is possible only as long as the pills are available to be counted. Participants sometimes forget or neglect to bring their pills to the clinic to be counted. In such circumstances, the investigator may ask the participant to count the pills himself at home and to notify the investigator of the result by telephone. Obviously, these data may be less reliable. The frequency with which data on pill counts are missing gives an estimate of the reliability of pill count as an adherence measure.

In monitoring pill count, the investigators ought to anticipate questions of interest to readers of the trial report when published. What was the overall adherence to the protocol prescription? If the overall adherence with the intervention was reduced, what was the main reason for the reduction? Were the participants prescribed a reduced dose of the study medication, or did they not follow the investigator's prescription? Was it because of intervening life events, specific side effects or was it simply forgetfulness? The answers to these questions may increase the understanding and interpretation of the results of the trial.

When discussing adherence assessed by pill count, the investigator has to keep in mind that these data may be inflated and misleading. Additionally, these data do not include information from participants who miss a visit. Most participants tend to overestimate their adherence either in an effort to please the investigator or because of faulty memory. Those who miss one or more visits typically have low adherence. Therefore, the adherence data should be viewed within the framework of all participants who are scheduled to be seen at a particular visit. There is general agreement on one point – the participant who says he did not take his study medication can be trusted.

Electronic monitoring of adherence has been used [28]. A device electronically records drug package opening times and duration, thus, describes dosing histories. The correlation between package openings and measured drug concentrations in serum is very high. The obvious advantage of electronic monitoring is that the dose-timing can be assessed to see if it is punctual and regular. In an HIV trial, overall adherence was 95%, but only 81% of the doses were taken within the prescribed dosing interval (±3 h) [29]. In a study of hypertensive participants, about 10% of the scheduled doses were omitted on any day [30]. Drug holidays, defined as omissions of all doses during three or more days, were recorded in 43% of the participants. An interesting observation was that participants with dosing problems were more likely later to become permanent drop-outs. It is not known whether or to what extent low adherence to dose-timing influences the trial findings.

Indirect information on adherence can also be obtained through interviews or record keeping by the participant. A diet study might use a 24-hour recall or a 7-day food record. Exercise studies may use diaries to record frequency and kind of exercise. Trials of people with angina might record frequency of attacks or pain and nitroglycerin consumption.

There are two major direct methods for measuring adherence. *Biochemical analyses* are sometimes made on either blood or urine in order to detect the presence of the active drug or metabolites. A limitation in measuring substances in urine or blood is the short half-life of most drugs. Therefore, laboratory determinations usually indicate only what has happened in the preceding day or two. A control participant who takes the active drug (obtained from a source outside the trial) until the day prior to a clinic visit, or a participant in the intervention group who takes the active drug only on the day of the visit might not always be detected as being a poor adherer. Moreover, drug adherence in participants taking an inert placebo tablet cannot be assessed by any laboratory determination. Adding a specific chemical substance such as riboflavin can serve as a marker in cases where the placebo, the drug or its metabolites are difficult to measure. However, the same drawbacks apply to markers as to masking substances – the risk of toxicity in long-term use may outweigh benefits.

Laboratory tests obtained on occasions not associated with clinic visits may give a better picture of regular or true adherence. Thus, the participant may be instructed, at certain intervals, to send a vial of urine to the clinic. Such a technique is of value only so long as the participant does not associate this request with an adherence monitoring procedure. In at least one study, information obtained in this manner contributed no additional information to laboratory results done at scheduled visits, except perhaps as a confirmation of such results.

Measurement of *physiological response variables* can be helpful in assessing adherence. Cholesterol reduction by drug or diet is unlikely to occur in 1 or 2 days. Therefore, a participant in the intervention group cannot suddenly adhere with the regimen the day before a clinic visit and expect to go undetected. Similarly, the cholesterol level of a nonadherent control participant is unlikely to rise in the 1 day before a visit if he skips the nonstudy lipid-lowering drug. Other physiological response variables that might be monitored are blood pressure in an antihypertensive study, carbon monoxide in a smoking study, platelet aggregation in an aspirin study, and graded exercise in an

exercise study. In all these cases, the indicated response variable would not be the primary response variable but merely an intermediate indicator of adherence to the intervention regimen. Unfortunately, not every person responds in the same way to medication, and some measures, such as triglyceride levels, are highly variable. Therefore, indications of low adherence of individual participants using these measures are not easily interpreted. Group data, however, may be useful.

Another aspect of monitoring deals with participant adherence to study procedures such as attendance at scheduled visits or *visit adherence*. One of the major purposes of these visits is to collect response variable data. The data will be better if they are more complete. Thus, completeness of data in itself can be a measure of the quality of a clinical trial. Studies with even a moderate amount of missing data or participants lost to follow-up could give misleading results and should be interpreted with caution. By reviewing the reasons why participants missed scheduled clinic visits, the investigator can identify factors that can be corrected or improved. Having the participants come in for study visits facilitates and encourages adherence to study medication. Study drugs are dispensed at these visits, and the dose is adjusted when necessary.

From a statistical viewpoint, every randomized participant should be included in the analysis (Chaps. 8 and 17). Consequently, the investigator must keep trying to get all participants back for scheduled visits until the trial is over. Even if a participant is taken off the study medication by an investigator or stops taking it, he should be encouraged to come in for regular study visits. Complete follow-up data on the response variables are critical so that visit adherence is important. In addition, participants do change their minds. For a long time, they may want to have nothing to do with the trial and later may agree to come back for visits and even resume taking their assigned intervention regimen. Special attention to the specific problems of each participant withdrawn from the trial and an emphasis on potential contribution to the trial can lead to successful retrieval of a large proportion of withdrawn participants. Inasmuch as the participant will be counted in the analysis, leaving open the option for the participant to return to active participation in the study is worthwhile.

Dealing with Low Adherence

A commonly asked question is whether a low adherence rate should be discussed directly with study participants. There is a consensus that any discussion should not be confrontational. The preferred approach is to open any discussion by saying that adherence to medications for many people can be very difficult. After being given examples of common reasons for low adherence, many participants seem to be more willing to discuss their own situations and adherence problems. Thus, sympathy and understanding may be helpful if followed by specific recommendations regarding ways to improve adherence. A large number of interviewing techniques of patients in the clinical setting are discussed by Shea [31].

If low adherence is related to difficulties making appointments, it may be useful to offer more convenient clinic hours, such as evenings and weekends as mentioned

above. Home visits are another option for participants with disabilities who have difficulties making it to the clinic. For participants who have moved, the investigator might be able to arrange for follow-up in other cities.

A remarkable recovery program was developed and implemented by Probstfield et al. [32]. Through participant counseling, the investigators succeeded in about 90% of the 36 drop-outs in approximately 6 months to return for clinic visits. Even more notable was the virtual absence of recidivism over the remaining 5 years of intervention. Approximately 70% of the drop-outs resumed taking their study medication, though typically at a lower dose than specified in the protocol.

One of the challenges in clinical trials is the complete ascertainment of response variables in participants who are no longer actively involved in the trial. The internet provides opportunities to track participants lost to follow-up. There are both fee-for-service and free search engines. The basic information required for a search is complete name, birth date, and Social Security Number or other specific identification number. These searches are more effective if several and different search engines are employed.

Steps should be taken to prevent situations in which participants request that they never be contacted. These are sometimes referred to as complete withdrawal. Participants who end their active participation in a clinical trial often agree to be contacted at the end of the trial for ascertainment of key response variables. For those who are lost to follow-up, but have not withdrawn their consent, alternative sources of information are family members and medical providers. The goal is to limit the amount of missing information.

Special Populations

Although the approaches to dealing with prevention of low adherence and maintenance of high adherence are applicable to people in general, there are factors that need consideration when dealing with special populations. Elderly people represent a growing number of participants in clinical trials. There is a rich literature on factors that may influence adherence and on strategies to increase adherence in the clinical setting among older people. Many of these are highly relevant for clinical trials. The motivation to adhere to an intervention can be difficult to promote in persons who are not fully functional. Since metabolism and physiology change with age, finding the proper dose of an intervention in elderly subjects represents another challenge. Polypharmacy and sometimes complex or inadequate instructions can lead to failure to take the study medication as prescribed. Elderly participants typically have more health complaints than their younger counterparts. Drug interactions are a concern that applies even to over-the-counter drugs. Assessment of intervention-related adverse reactions is typically difficult.

In addition to the elderly, there are many other groups that require special attention, either for physical, mental, or cultural reasons, and investigators need to be aware of those needs, and access them as best they can [33–35].

References

1. Haynes RB, Taylor DW, Sackett DL (eds.). *Compliance in Health Care.* Baltimore: Johns Hopkins University Press, 1979.
2. Gossec L, Tubach F, Dougados M, Ravaud P. Reporting of adherence to medication in recent randomized controlled trials of 6 chronic diseases: A systematic literature review. *Am J Med Sci* 2007;334:248–254.
3. Peterson AM, Takiya L, Finley R. Meta-analysis of trials of interventions to improve medication adherence. *Am J Health Syst Pharm* 2003;60:657–665.
4. Boden WE, O'Rourke RA, Teo KK, et al. for the COURAGE Trial Research Group. Optimal medical therapy with or without PCI for stable coronary disease. *N Engl J Med* 2007;356:1503–1516.
5. DiMatteo MR. Variations in patients' adherence to medical recommendations. A quantitative review of 50 years of research. *Med Care* 2004;42:200–209.
6. Shumaker SA, Ockene JK, Riekert KA (eds.). *The Handbook of Health Behavior Change* (3rd edition). New York: Springer Publishing Company, 2009.
7. Osterberg L, Blaschke T. Adherence to medication. *N Engl J Med* 2005;353:487–497.
8. Härkäpää K, Järvikoski A, Mellin G, Hurri H. A controlled study of the outcome of inpatient and outpatient treatment of low back pain. Part I. Pain, disability, compliance, and reported treatment benefits three months after treatment. *Scand J Rehabil Med* 1989;21:81–89.
9. McLean N, Griffin S, Toney K, Hardeman W. Family involvement in weight control, weight maintenance and weight-loss interventions: A systematic review of randomised trials. *Intern J Obesity* 2003;27:987–1005.
10. Voils CI, Yancy Jr WS, Kovac S, et al. Study protocol: Couples Partnering for Lipid Enhancing Strategies (CouPLES) – a randomized, controlled trial. *Trials* 2009;10:10–19.
11. Sacks FM, Svetkey LP, Vollmer WM, et al. Effects on blood pressure of reduced dietary sodium and the Dietary Approaches to Stop Hypertension (DASH) diet. *N Engl J Med* 2001;344:3–10.
12. Claxton AJ, Cramer J, Pierce C. A systematic review of the associations between dose regimens and medication compliance. *Clin Ther* 2001;23:1296–1310.
13. Probstfield JL. The clinical trial prerandomization compliance (adherence) screen. In Cramer JA, Spilker B (eds.). *Patient Compliance in Medical Practice and Clinical Trials.* New York: Raven Press, 1991.
14. Berger VW, Rezvani A, Makarewicz VA. Direct effect on validity of response run-in selection in clinical trials. *Control Clin Trials* 2003;24:156–166.
15. Lang JM, Buring JE, Rosner B, et al. Estimating the effect of the run-in on the power of the Physicians' Health Study. *Stat Med* 1991;10:1585–1593.
16. Lee S, Walker JR, Jakul L, Sexton K. Does elimination of placebo responders in a placebo run-in increase the treatment effect in randomized clinical trials? A meta-analytic evaluation. *Depress Anxiety* 2004;19:10–19.
17. Kingry C, Bastien A, Booth G, et al. for the ACCORD Study Group. Recruitment strategies in the Action to Control Cardiovascular Risk of Diabetes (ACCORD) Trial. *Am J Cardiol* 2007;99[suppl]:68i–79i.
18. Dunbar-Jacob J, Gemmel LA, Schlenk EA. Predictors of patient adherence: Patient characteristics. In Shumaker SA, Ockene JK, Riekert KA (eds.). *The Handbook of Health Behavior Change* (3rd edition). New York: Springer Publishing Company, 2009, pp. 397–410.
19. Williams SL, DiMatteo MR, Haskard KB. Psychological barriers to adherence and lifestyle change. In Shumaker SA, Ockene JK, Riekert KA (eds.). *The Handbook of Health Behavior Change* (3rd edition). New York: Springer Publishing Company, 2009, pp. 445–461.
20. Murdaugh CL, Insel K. Problems with adherence in the elderly. In Shumaker SA, Ockene JK, Riekert KA (eds.). *The Handbook of Health Behavior Change* (3rd edition). New York: Springer Publishing Company, 2009, pp. 499–518.

21. Franklin PD, Farzanfar R, Thompson DD. E-health strategies to support adherence. In Shumaker SA, Ockene JK, Riekert KA (eds.). *The Handbook of Health Behavior Change* (3rd edition). New York: Springer Publishing Company, 2009, pp. 169–190.
22. Schillinger D, Piette J, Grumbach K, et al. Closing the loop. Physician communication with diabetic patients who have low health literacy. *Arch Intern Med* 2003;163:83–90.
23. Barnes K. Patients provide insight into trial participation. Outsourcing-Pharma.com, July 4, 2007. www.outsourcing-pharma.com/content/view/print/135930.
24. Dexheimer JW, Sanders DL, Rosenbloom ST, et al. Prompting clinicians: A systematic review of prevention care reminders. *AMIA Annu Symp Proc* 2005;2005:938.
25. Kripalani S, Yao X, Haynes RB. Interventions to enhance medication adherence in chronic medical conditions. *Arch Intern Med* 2007;167:540–550.
26. Haynes RB, Ackloo E, Sahota N, et al. Interventions for enhancing medication adherence (Review). *Cochrane Database Syst Rev* 2009, Issue 2.
27. Guthrie RM. The effects of postal and telephone reminders on compliance with pravastatin therapy in a national registry: Results of the first myocardial infarction risk reduction program. *Clin Ther* 2001;23:970–980.
28. Otsuki M, Clerisme-Beaty E, Rand CS, Riekert KA. Measuring adherence to medication regimens in clinical care and research. In Shumaker SA, Ockene JK, Riekert KA (eds.). *The Handbook of Health Behavior Change* (3rd edition). New York: Springer Publishing Company, 2009, pp. 309–325.
29. Vrijens B, Rousset E, Rode R, et al. Successful projection of the time course of drug concentration in plasma during a 1-year period from electronically compiled dosing-time data used as input to individually parameterized pharmacokinetic models. *J Clin Pharmacol* 2005;45:461–467.
30. Vrijens B, Vincze G, Kristanto P, et al. Adherence to prescribed antihypertensive drug treatments: Longitudinal study of electronically compiled dosing histories. *Br Med J* 2008;336:1114–1117.
31. Shea SC. *Improving Medication Adherence: How to Talk with Patients About Their Medications.* Philadelphia: Lippincott, Williams & Wilkins, 2006.
32. Probstfield JL, Russell ML, Henske JC, et al. Successful program for recovery of dropouts to a clinical trial. *Am J Med* 1986;80:777–784.
33. Whitt-Glover MC, Beech BM, Bell RA, et al. Health disparities and minority health. In Shumaker SA, Ockene JK, Riekert KA (eds.). *The Handbook of Health Behavior Change* (3rd edition). New York: Springer Publishing Company, 2009, pp. 589–606.
34. Ievers-Landis CE, Witherspoon D. Lifestyle interventions for the young. In Shumaker SA, Ockene JK, Riekert KA (eds.). *The Handbook of Health Behavior Change* (3rd edition). New York: Springer Publishing Company, 2009, pp. 483–498.
35. Rapoff MA. Adherence issues among adolescents with chronic diseases. In Shumaker SA, Ockene JK, Riekert KA (eds.). *The Handbook of Health Behavior Change* (3rd edition). New York: Springer Publishing Company, 2009, pp. 545–588.

Chapter 15
Survival Analysis

This chapter reviews some of the fundamental concepts and basic methods in survival analysis. Frequently, event rates such as mortality or occurrence of nonfatal myocardial infarction are selected as primary response variables. The analysis of such event rates in two groups could employ the chi-square statistic or the equivalent normal statistic for the comparison of two proportions. However, when the length of observation is different for each participant, estimating an event rate is more complicated. Furthermore, simple comparison of event rates between two groups is not necessarily the most informative type of analysis. For example, the 5-year survival for two groups may be nearly identical, but the survival rates may be quite different at various times during the 5 years. This is illustrated by the survival curves in Fig. 15.1. This figure shows survival probability on the vertical axis and time on the horizontal axis. For Group *A*, the survival rate (or 1 – the mortality rate) declines steadily over the 5 years of observation. For Group *B*, however, the decline in the survival rate is rapid during the first year and then levels off. Obviously, the survival experience of the two groups is not the same although the mortality rate at 5 years is nearly the same. If only the 5-year survival rate is considered, Group *A* and Group *B* appear equivalent. Curves like these might reasonably be expected in a trial of surgical versus medical intervention, where surgery might carry a high initial operative mortality.

Fundamental Point

Survival analysis methods are important in trials where participants are entered over a period of time and have various lengths of follow-up. These methods permit the comparison of the entire survival experience during the follow-up and may be used for the analysis of time to any dichotomous response variable such as a nonfatal event or an adverse event.

A review of the basic techniques of survival analysis can be found in elementary statistical textbooks [1–6] as well as in overview papers [7]. A more complete and

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_15, © Springer Science+Business Media, LLC 2010

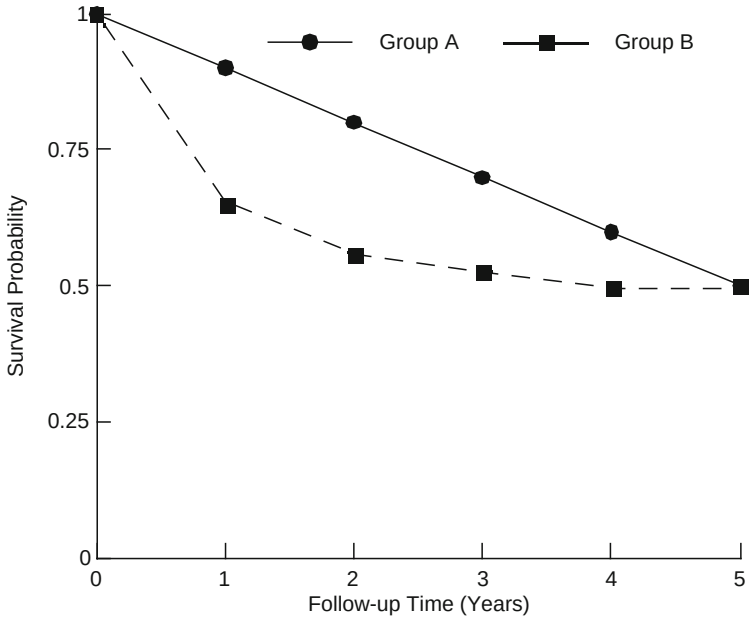


Fig. 15.1 Survival experience for two groups (*A* and *B*)

technical review is in other texts [8–11]. Many methodological advances in the field have occurred, and this book will not be able to cover all developments. The following discussion will concern two basic aspects: first, estimation of the survival experience or survival curve for a group of participants in a clinical trial and second, comparison of two survival curves to test whether the survival experience is significantly different. Although the term survival analysis is used, the methods are more widely applicable than to just survival. The methods can be used for any dichotomous response variable when the time from enrollment to the time of the event, not just the fact of its occurrence, is an important consideration. For ease of communication, we shall use the term event, unless death is specifically the event.

Estimation of the Survival Curve

The graphical presentation of the total survival experience during the period of observation is called the survival curve, and the tabular presentation is called the lifetable. In the sample size discussion (Chap. 8), we utilized a parametric model to represent a survival curve, denoted $S(t)$, where t is the time of follow-up. A classic parametric form for $S(t)$ is to assume an exponential distribution $S(t) = e^{-\lambda t} = \exp(-\lambda t)$, where λ is the hazard rate [9]. If we estimate λ, we have an estimate for $S(t)$. One possible estimate for the hazard ratio is the number of observed events divided by the total exposure time of the person at risk of the event. Other estimates are also

available and are described later. While this estimate is not difficult to obtain, the hazard rate may not be constant during the trial. If λ is not constant, but rather a function of time, we can define a hazard rate $\lambda(t)$, but now the definition is more complicated. Specifically, $S(t) = \exp\left[\int_0^t \lambda(s)ds\right]$, that is, the exponential of the area under the hazard function curve from time 0 to time t. Furthermore, we cannot always be guaranteed that the observed survival data will be described well by the exponential model, even though we often make this assumption for computing sample size. Thus, biostatisticians have relied on parameter-free or non-parametric ways to estimate the survival curve.

This chapter will cover two similar non-parametric methods, the Kaplan–Meier method [12] and the Cutler–Ederer method [13] for estimating the true survival curve or the corresponding lifetable. Before a review of these specific methods, however, it is necessary to explain how the survival experience is typically obtained in a clinical trial and to define some of the associated terminology.

The clinical trial design may, in a simple case, require that all participants be observed for T years. This is referred to as the follow-up or exposure time. If all participants are entered as a single cohort at the same time, the actual period of follow-up is the same for all participants. If, however, as in most clinical trials, the entry of participants is staggered over some recruitment period, then equal periods of follow-up may occur at different calendar times for each participant, as illustrated in Fig. 15.2.

A participant may have a study event during the course of follow-up. The event time is the accumulated time from entry into the study to the event. The interest is not in the actual calendar date when the event took place but rather the interval of time from entry into the trial until the event. Figures 15.3 and 15.4 illustrate the way the actual survival experience for staggered entry of participants is translated for the analysis. In Fig. 15.3, participants 2 and 4 had an event while participants 1 and 3

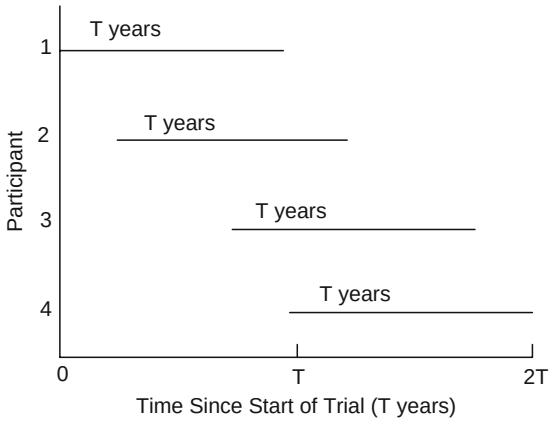


Fig. 15.2 *T* year follow-up time for four participants with staggered entry

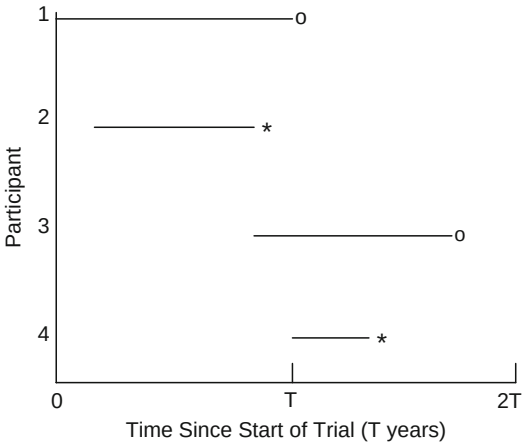


Fig. 15.3 Follow-up experience of four participants with staggered entry: two participants with observed events (*asterisks*) and two participants followed for time *T* without events (*open circles*)

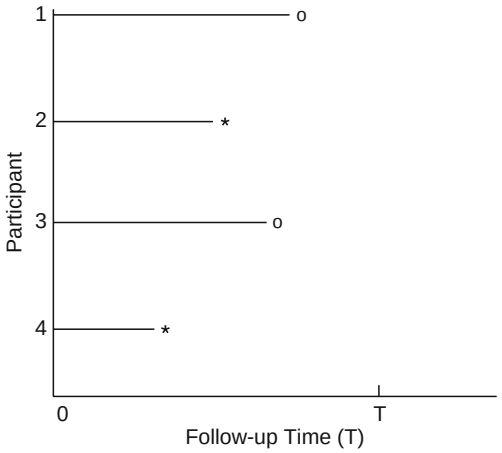


Fig. 15.4 Follow-up experience of four participants with staggered entry converted to a common starting time: two participants with observed events (*asterisks*) and two participants followed for time *T* without events (*open circle*)

did not during the follow-up time. Since, for each participant, only the time interval from entry to the end of the scheduled follow-up period or until an event is of interest, the time of entry can be considered as time zero for each participant. Figure 15.4 illustrates the same survival experience as Fig. 15.3, but the time of entry is considered as time zero.

Some participants may not experience an event before the end of observation. The follow-uptime or exposure time for these participants is said to be *censored*;

that is, the investigator does not know what happened to these participants after they stopped participating in the trial. Another example of censoring is when participants are entered in a staggered fashion, and the study is terminated at a common date before all participants have had at least their complete *T* years of follow-up. Later posttrial events from these participants are also unobserved, but the reason for censoring is administrative. Administrative censoring could also occur if a trial is terminated prior to the scheduled time because of early benefits or harmful effects of the intervention. In these cases, censoring is assumed to be independent of occurrence of events.

Figure 15.5 illustrates several of the possibilities for observations during follow-up. Note that in this example, the investigator has planned to follow all participants to a common termination time, with each participant being followed for at least *T* years. The first three participants were randomized at the start of the study. The first participant was observed for the entire duration of the trial with no event, and her survival time was censored because of study termination. The second participant had an event before the end of follow-up. The third participant was lost to follow-up. The second group of three participants was randomized later during the course of the trial with experiences similar to the first group of three. Participants 7–11 were

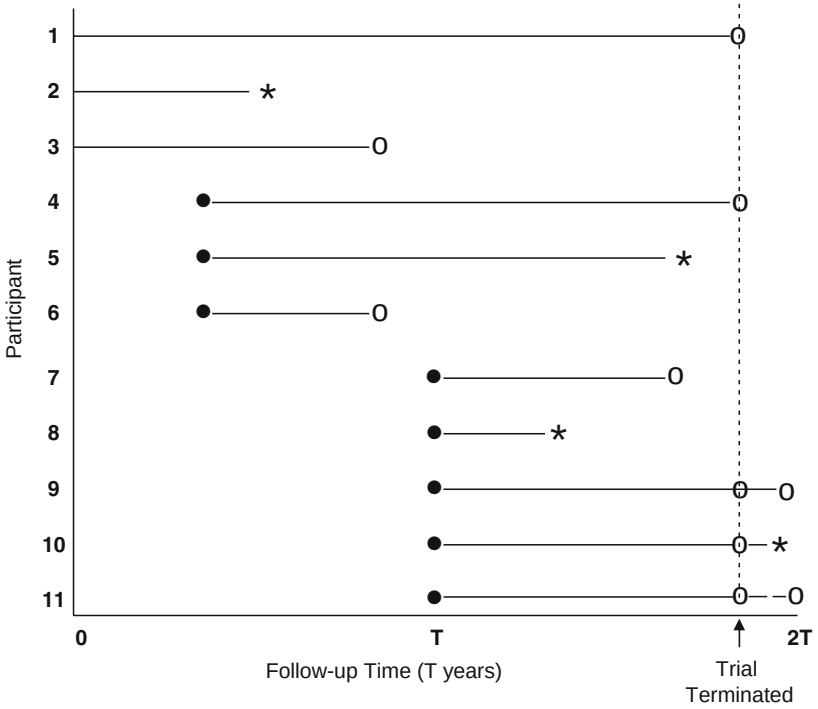


Fig. 15.5 Follow-up experience of 11 participants for staggered entry and a common termination time, with observed events (*asterisks*) censoring (*open circles*). Follow-up experience beyond the termination time is shown for participants 9–11

randomized late in the study and were not able to be followed for at least *T* years because the study was terminated early. Participant 7 was lost to follow-up and participant 8 had an event before *T* years of follow-up time had elapsed and before the study was terminated. Participant 9 was administratively censored but theoretically would have been lost to follow-up had the trial continued. Participant 10 was also censored because of early study termination, although she had an event afterwards which would have been observed had the trial continued to its scheduled end. Finally, the last participant who was censored would have survived for at least *T* years had the study lasted as long as first planned. The survival experiences illustrated in Fig. 15.5 would all be shifted to have a common starting time equal to zero as in Fig. 15.4. The follow-up time, or the time elapsed from calendar time of entry to calendar time of an event or to censoring could then be analyzed.

In summary then, the investigator needs to record for each participant the time of entry and the time of an event, the time of loss to follow-up, or whether the participant was still being followed without having had an event when the study is terminated. These data will allow the investigator to compute the survival curve.

Kaplan–Meier Estimate

In a clinical trial with staggered entry of participants and censored observations, survival data will be of varying degrees of completeness. As a very simple example, suppose that 100 participants were entered into a study and followed for 2 years. One year after the first group was started, a second group of 100 participants was entered and followed for the remaining year of the trial. Assuming no losses to follow-up, the results might be as shown in Table 15.1. For Group *I*, 20 participants died during the first year and of the 80 survivors, 20 more died during the second year. For Group *II*, which was followed for only 1 year, 25 participants died. Now suppose the investigator wants to estimate the 2-year survival rate. The only group of participants followed for 2 years was Group *I*. One estimate of 2-year survival, $p(2)$, would be $p(2) = 60/100$ or 0.60. Note that the first-year survival experience of Group *II* is ignored in this estimate. If the investigator wants to estimate 1 year

Table 15.1 Participants entered at two points in time (Group *I* and Group *II*) and followed to a common termination time[a]

Years of follow-up		Group *I*	Group *II*
1	Participants entered	100	100
	1st year deaths	20	25
	1st year survivors	80	75
2	Participants entered	80	
	2nd year deaths	20	
	2nd year survivors	60	

[a]After Kaplan and Meier [12]

survival rate, $p(1)$, she would observe that a total of 200 participants were followed for at least 1 year. Of those, 155 (80 + 75) survived the first year. Thus, $p(1) = 155/200$ or 0.775. If each group were evaluated separately, the survival rates would be 0.80 and 0.75. In estimating the 1-year survival rate, all the available information was used, but for the 2-year survival rate, the 1-year survival experience of Group *II* was ignored.

Another procedure for estimating survival rates is to use a conditional probability. For this example, the probability of 2-year survival, $p(2)$, is equal to the probability of 1-year survival, $p(1)$, times the probability of surviving the second year, given that the participant survived the first year, $p(2|1)$. That is, $p(2) = p(1)\ p(2|1)$. In this example, $p(1) = 0.775$. The estimate for $p(2|1)$ is $60/80 = 0.75$ since 60 of the 80 participants who survived the first year also survived the second year. Thus, the estimate for $p(2) = 0.775 \times 0.75$ or 0.58, which is slightly different from the previously calculated estimate of 0.60.

Kaplan and Meier [12] described how this conditional probability strategy could be used to estimate survival curves in clinical trials with censored observations. Their procedure is usually referred to as the Kaplan–Meier estimate, or sometimes the product-limit estimate, since the product of conditional probabilities leads to the survival estimate. This procedure assumes that the exact time of entry into the trial is known and that the exact time of the event or loss of follow-up is also known. For some applications, time to the nearest month may be sufficient, while for other applications the nearest day or hour may be necessary. Kaplan and Meier assumed that a death and loss of follow-up would not occur at the same time. If a death and a loss to follow-up are recorded as having occurred at the same time, this tie is broken on the assumption that the death occurred slightly before the loss to follow-up.

In this method, the follow-up period is divided into intervals of time so that no interval contains both deaths and losses. Let p_j be equal to the probability of surviving the *j*th interval, given that the participant has survived the previous interval. For the rest of this chapter, lower case *p* refers to the conditional probability of surviving a particular interval. Upper case *P* refers to the cumulative probability of surviving up through a specific interval. For intervals labeled *j* with deaths only,the estimate for p_j, which is $\hat{P}_j$, is equal to the number of participants alive at the beginning of the *j*th interval, n_j, minus those who died during the interval, δ_j, with this difference being divided by the number alive at the beginning of the interval, i.e., $\hat{P}_j = (n_j - \delta_j)/n_j$. For an interval *j* with only l_j losses, the estimate $\hat{P}_j$ is one. Such conditional probabilities for an interval with only losses would not alter the product. This means that an interval with only losses and no deaths may be combined with the previous interval.

Example: Suppose 20 participants are followed for a period of 1 year, and to the nearest tenth of a month, deaths were observed at the following times: 0.5, 1.5, 1.5, 3.0, 4.8, 6.2, 10.5 months. In addition, losses to follow-up were recorded at: 0.6, 2.0, 3.5, 4.0, 8.5, 9.0 months. It is convenient for illustrative purposes to list the deaths and losses together in ascending time with the losses indicated in parentheses. Thus, the following sequence is obtained: 0.5, (0.6), 1.5, 1.5, (2.0), 3.0, (3.5), (4.0),

Table 15.2 Kaplan–Meier lifetable for 20 participants followed for 1 year

Interval	Interval number	Time of death	n_j	δ_j	l_j	$\hat{P}_j$	$\hat{P}(t)$	Var $\hat{P}(t)$
[0.5,1,5)	1	0.5	20	1	1	0.95	0.95	0.0024
[1.5,3.0)	2	1.5	18	2	1	0.89	0.85	0.0068
[3.0,4.8)	3	3.0	15	1	2	0.93	0.79	0.0089
[4.8,6.2)	4	4.8	12	1	0	0.92	0.72	0.0114
[6.2,10.5)	5	6.2	11	1	2	0.91	0.66	0.0133
[10.5)	6	10.5	8	1	7[a]	0.88	0.58	0.0161

n_j: number of participants alive at the beginning of the jth interval
δ_j: number of participants who died during the jth interval
l_j: number of participants who were lost or censored during the jth interval
$\hat{P}_j$: estimate for p_j, the probability of surviving the jth interval given that the participant has survived the previous intervals
$\hat{P}(t)$: estimated survival curve
$V[\hat{P}(t)]$: variance of $\hat{P}(t)$
[a]Censored due to termination of study

4.8, 6.2, (8.5), (9.0), 10.5. The remaining seven participants were all censored at 12 months due to termination of the study.

Table 15.2 presents the survival experience for this example as a lifetable. Each row in the lifetable indicates the time at which a death or an event occurred. One or more deaths may have occurred at the same time, and they are included in the same row in the lifetable. In the interval between two consecutive times of death, losses to follow-up may have occurred. Hence, a row in the table actually represents an interval of time, beginning with the time of a death, up to but not including the time of the next death. In this case, the first interval is defined by the death at 0.5 months up to the time of the next death at 1.5 months. The columns labeled n_j, δ_j, and l_j correspond to the definitions given above and contain the information from the example. In the first interval, all 20 participants were initially at risk, and one died at 0.5 months; later in the interval (at 0.6 months), one participant was lost to follow-up. In the second interval, from 1.5 months up to 3.0 months, 18 participants were still at risk initially, two deaths were recorded at 1.5 months and one participant was lost at 2.0 months. The remaining intervals are defined similarly. The column labeled $\hat{P}_j$is the conditional probability of surviving the interval j and is computed as $(n_j - \delta_j)/n_j$ or $(20-1)/20 = 0.95$, $(18-2)/18 = 0.89$, etc. The column labeled (t) is the estimated survival curve and is computed as the accumulated product of the p_j's ($0.85 = 0.95 \times 0.89$, $0.79 = 0.95 \times 0.89 \times 0.93$, etc.).

The graphical display of the next to last column of Table 15.2, $\hat{P}(t)$, is given in Fig. 15.6. The step function appearance of the graph is because the estimate of $P(t)$, $\hat{P}(t)$ is constant during an interval and changes only at the time of a death. With very large sample sizes and more observed deaths, the step function has smaller steps and looks more like the usually visualized smooth survival curve. If no censoring occurs, this method simplifies to the number of survivors divided by the total number of participants who entered the trial.

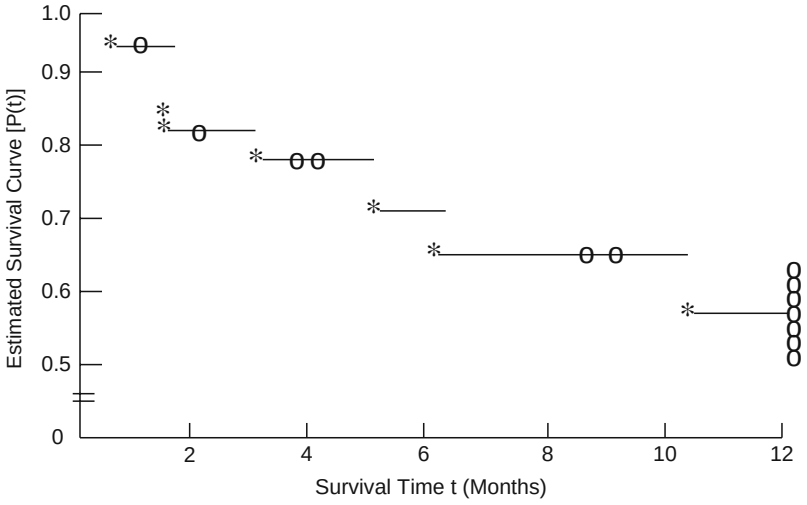


Fig. 15.6 Kaplan–Meier estimate of a survival curve, $\hat{P}(t)$, from a 1-year study of 20 participants, with observed events (*asterisks*) and censoring (*open circles*)

Because $\hat{P}(t)$ is an estimate of $P(t)$, the true survival curve, the estimate will have some variation due to the sample selected. Greenwood [14] derived a formula for estimating the variance of an estimated survival function which is applicable to the Kaplan–Meier method. The formula for the variance of (t), denoted $V[\hat{P}(t)]$ is given by

$$V[\hat{P}(t)] = \hat{P}^2(t)\sum_{j=1}^{\kappa} \frac{\delta_j}{n_j(n_j - \delta_j)}$$

where n_j and δ_j are defined as before, and K is the number of intervals. In Table 15.2, the last column labeled $V[\hat{P}(t)]$ represents the estimated variances for the estimates of $P(t)$ during the six intervals. Note that the variance increases as one moves down the column. When fewer participants are at risk, the ability to estimate the survival experience is diminished.

Other examples of this procedure, as well as a more detailed discussion of some of the statistical properties of this estimate, are provided by Kaplan and Meier [12]. Computer programs are available [15] so that survival curves can be obtained quickly, even for very large sets of data.

The Kaplan–Meier curve can also be used to estimate the hazard rate, λ, if the survival curve is exponential. For example, if the median survival time is estimated as T_M, then $0.5 = S(T_M) = e^{-\lambda T_M} = \exp(-\lambda T_M)$ and thus $\hat{\lambda} = \ln(05)/T_M$ as an estimate of λ. Then the estimate for $S(t)$ would be $\exp(-\hat{\lambda}t)$. In comparison to the Kaplan–Meier, another parametric estimate for $S(t)$ at time t_j, described by Nelson [16], is

$$\hat{S}(t_j) = \exp\left\{-\sum_{i=1}^{j} \delta_i / n_j\right\}$$

where δ_i is the number of events in the ith interval and n_i is the number at risk for the event. While this is a straightforward estimate, the Kaplan–Meier does not assume an underlying exponential distribution and thus is used more than this type of estimator.

Cutler–Ederer Estimate

In the Kaplan–Meier estimate, it is required that the exact time of death or loss be known so that the observations could be ranked, or at least grouped appropriately, into intervals with deaths preceding losses. For some studies, all that is known is that within an interval of time from t_{j-1} to t_j, denoted $(t_{j-1} - t_j)$, δ_j deaths and l_j losses occurred among the n_j participants at risk. Within that interval, the order in which the events and losses occurred is unknown. In the Kaplan–Meier procedure, the intervals were chosen so that all deaths preceded all losses in any interval.

In the Cutler–Ederer or actuarial estimate [13], the assumption is that the deaths and losses are uniformly distributed over an interval. On the average, this means that one half the losses will occur during the first half of the interval. The estimate for the probability of surviving the jth interval, given that the previous intervals were survived, is $\hat{P}_j$, where

$$\hat{P}_j = \frac{n_j - \delta_j - 0.5\lambda_j}{n_j - 0.5\lambda_j}$$

Notice the similarity to the Kaplan–Meier definition. The modification is that the λ_j losses are assumed to be at risk, on the average, one half the time and thus should be counted as such. These conditional probabilities, $\hat{P}_j$, are then multiplied together as in the Kaplan–Meier procedure to obtain an estimate, $\hat{P}(t)$, of the survival function at time t. The estimated variance for $\hat{P}(t)$ in this case is given by

$$V[\hat{P}(t)] = \hat{P}^2(t) \sum_{j=1}^{\kappa} \frac{\delta_j}{(n_j - 0.5\lambda_j)(n_j - 0.5\lambda_j - \delta_j)}$$

Specific applications of this method are described by Cutler and Ederer [13]. The parallel to the example shown in Table 15.2 would require recomputing the $\hat{P}_j$, $\hat{P}(t)$ and $V[\hat{P}(t)]$.

Comparison of Two Survival Curves

We have just discussed how to estimate the survival curve in a clinical trial for a single group. For two groups, the survival curve would be estimated for each group separately. The question is whether the two survival curves $P_C(t)$ and $P_I(t)$, for the control and intervention groups respectively, are different based on the estimates $\hat{P}_C(t)$ and $\hat{P}_I(t)$.

Point-by-Point Comparison

One possible comparison between groups is to specify a time t^* for which survival estimates have been computed using the Kaplan–Meier [12] or Cutler–Ederer [13] method. At time t^*, one can compare the survival estimates $\hat{P}_C(t^*)$ and $\hat{P}_I(t^*)$ using the statistic

$$Z(t^*) = \frac{\hat{P}_C(t^*) - \hat{P}_I(t^*)}{\{V[\hat{P}_C(t^*)] + V[\hat{P}_I(t^*)]\}^{1/2}}$$

where $V[\hat{P}_C(t^*)]$ and $V[\hat{P}_I(t^*)]$ are the Greenwood estimates of variance [14]. The statistic $Z(t^*)$ has approximately a normal distribution with mean zero and variance one under the null hypothesis that $\hat{P}_C(t^*) = \hat{P}_I(t^*)$. The problem with this approach is the multiple looks issue described in Chap. 16. Another problem exists in interpretation. For example, what conclusions should be drawn if two survival curves are judged significantly different at time t^* but not at any other points? The issue then becomes, what point in the survival curve is most important.

For some studies with a T year follow-up, the T year mortality rates are considered important and should be tested in the manner just suggested. Annual rates might also be considered important and therefore, compared. One criticism of this suggestion is that the specific points may have been selected post hoc to yield the largest difference based on the observed data. One can easily visualize two survival curves for which significant differences are found at a few points. However, when survival curves are compared, the large differences indicated by these few points are not supported by the overall survival experience. Therefore, point-by-point comparisons are not recommended unless a few points can be justified prior to data analysis and are specified in the protocol.

Comparison of Median Survival Times

One summary measure of survival experience is the time at which 50% of the cohort has had the event. One common and easy way to estimate the median survival time is from the Kaplan–Meier curve. (See for example, Altman [4].)

This assumes that the cohort has been followed long enough so that over one-half of the individuals have had the event. Confidence intervals may be computed for the median survival times [17]. If this is the case, we can compare the median survival times for intervention and control M_I and M_C, respectively. This is most easily done by estimating the ratio of the estimates M_I/M_C. A ratio larger than unity implies that the intervention group has a larger median survival and thus a better survival experience. A ratio less than unity would indicate the opposite.

We can estimate 95% confidence intervals for M_I/M_C by

$$(M_I/M_C)\mathrm{e}^{-1.96S}, (M_I/M_C)e^{+1.96S}$$

where the standard deviation, S, of M_I/M_C is computed as

$$S = \sqrt{1/(O_I + O_S)}$$

for cases where the survival curves are approximately exponential, and O_I = the total number of events in the intervention group (i.e., $\sum \delta_i$) and O_C = the total number of events in the control group.

Total Curve Comparison

Because of the limitations of comparison of point-by-point estimates, Gehan [18] and Mantel [19] originally proposed statistical methods to assess the overall survival experience. These two methods were important steps in the development of analytical methods for survival data. They both assume that the hypothesis being tested is whether two survival curves are equal, or whether one is consistently different from the other. If the two survival curves cross, these methods should be interpreted cautiously. Since these two original methods, an enormous literature has developed on comparison of survival curves and is summarized in several texts [8–11]. The basic methods described here provide the fundamental concepts used in survival analysis.

Mantel [19] proposed the use of the procedure described by Cochran [20] and Mantel and Haenszel [21] for combining a series of 2×2 tables. In this procedure, each time, t_j, a death occurs in either group, a 2×2 table is formed as follows:

The entry a_j represents the observed number of deaths at time t_j in the intervention group, and c_j represents the observed number of deaths at time t in the control group.

	Death at time t_j	Survivors at time t_j	At risk prior to time t_j
Intervention	a_j	b_j	$a_j + b_j$
Control	c_j	d_j	$c_j + d_j$
	$a_j + c_j$	$b_j + d_j$	n_j

At least a_j or c_j must be non-zero. One could create a table at other time periods (that is, when a_j and c_j are zero), but this table would not make any contribution to the statistic. Of the n_j participants at risk just prior to time t_j, a_j+b_j were in the intervention group and c_j+d_j were in the control group. The expected number of deaths in the intervention group, denoted $E(a_j)$, can be shown to be

$$E(a_j) = (a_j + c_j)(a_j + b_j) / n_j$$

and the variance of the observed number of deaths in the intervention group, denoted as $V(a_j)$ is given by

$$V(a_j) = \frac{(a_j + c_j)(b_j + d_j)(a_j + b_j)(c_j + d_j)}{n_j^2(n_j - 1)}$$

These expressions are the same as those given for combining 2×2 tables in the Appendix of Chap. 17. The Mantel–Haenszel (MH) statistic is given by

$$\text{MH} = \left\{\sum_{j=1}^{\kappa} a_j - E(a_j)\right\}^2 \Big/ \sum_{j=1}^{\kappa} V(a_j)$$

and has approximately a chi-square distribution with one degree of freedom, where K is the number of distinct event times in the combined intervention and control groups. The square root of MH, $Z_{\text{MH}} = \sqrt{\text{MH}}$, has asymptotically a standard normal distribution [22, 23].

Application of this procedure is straightforward. First, the times of events and losses in both groups are ranked in ascending order. Second, the time of each event, and the total number of participants in each group who were at risk just before the death (a_j+b_j, c_j+d_j) as well as the number of events in each group (a_j, c_j) are determined. With this information, the appropriate 2×2 tables can be formed.

Example: Assume that the data in the example shown in Table 15.2 represent the data from the control group. Among the 20 participants in the intervention group, two deaths occurred at 1.0 and 4.5 months with losses at 1.6, 2.4, 4.2, 5.8, 7.0, and 11.0 months. The observations, with parentheses indicating losses, can be summarized as follows:

Intervention: 1.0, (1.6), (2.4), (4.2), 4.5, (5.8), (7.0), (11.0)
Control: 0.5, (0.6), 1.5, 1.5, (2.0), 3.0, (3.5), (4.0), 4.8, 6.2, (8.5), (9.0), 10.5.

Using the data described above, with remaining observations being censored at 12 months, Table 15.3 shows the eight distinct times of death, (t_j), the number in each group at risk prior to the death, ($a_j + b_j$, $c_j + d_j$), the number of deaths at time t_j, (a_j, c_j), and the number of participants lost to follow-up in the subsequent interval (l_j). The entries in this table are similar to those given for the Kaplan–Meier life-table shown in Table 15.2. Note in Table 15.3, however, that the observations from two groups have been combined with the net result being more intervals.

Table 15.3 Comparison of survival data for a control group and an intervention group using the Mantel-Haenszel procedures

Rank	Event times	Intervention			Control			Total	
j	t_j	$a_j + b_j$	a_j	l_j	$c_j + d_j$	c_j	l_j	$a_j + c_j$	$b_j + d_j$
1	0.5	20	0	0	20	1	1	1	39
2	1.0	20	1	0	18	0	0	1	37
3	1.5	19	0	2	18	2	1	2	35
4	3.0	17	0	1	15	1	2	1	31
5	4.5	16	1	0	12	0	0	1	27
6	4.8	15	0	1	12	1	0	1	26
7	6.2	14	0	1	11	1	2	1	24
8	10.5	13	0	13	8	1	7	1	20

$a_j + b_j$ = number of participants at risk in the intervention group prior to the death at time t_j
$c_j + d_j$ = number of participants at risk in the control group prior to the death at time t_j
a_j = number of participants in the intervention group who died at time t_j
c_j = number of participants in the control group who died at time t_j
l_j = number of participants who were lost or censored between time t_j and t_{j+1}
$a_j + c_j$ = number of participants in both groups who died at time t_j
$b_j + d_j$= number of participants in both group who are alive minus the number who died at time t_j

The entries in Table 15.3 labeled a_j+b_j, c_j+d_j, a_j+c_j, and b_j+d_j become the entries in the eight 2×2 tables shown in Table 15.4.

The Mantel–Haenszel statistic can be computed from these eight 2×2 tables (Table 15.4) or directly from Table 15.3. The term $\sum_{j=1}^{8} a_j = 2$ since there are only two deaths in the intervention group. Evaluation of the term $\sum_{j=1}^{8} E(a_j) = 20/40 + 20/38 + 2\times19/37 + 17/32 + 16/28 + 15/27 + 14/25 + 13/21$ or $\sum_{j=1}^{8} E(a_j) = 4.89$. The value for $\sum_{j=1}^{8} V(a_j)$ is computed as

$$\sum_{j=1}^{8} V(a_j) = \frac{(1)(39)(20)(20)}{(40)^2(39)} + \frac{(1)(37)(20)(18)}{(38)^2(37)} + \ldots$$

This term is equal to 2.21. The computed statistic is $MH = (2-4.89)^2/2.21 = 3.78$. This is not significant at the 0.05 significance level for a chi-square statistic with one degree of freedom. The MH statistic can also be used when the precise time of death is unknown. If death is known to have occurred within an interval, 2×2 tables can be created for each interval and the method applied. For small samples, a continuity correction is sometimes used. The modified numerator is

$$\left\{\left|\sum_{j=1}^{\kappa}[a_j - E(a_j)]\right| - 0.5\right\}^2$$

Table 15.4 Eight 2×2 tables corresponding to the event times used in the Mantel–Haenszel statistic in survival comparison of intervention (I) and control (C) groups

	D[b]	A[c]	R[d]
1. (0.5 mo)[a]			
I	0	20	20
C	1	19	20
	1	39	40
2. (1 mo)			
I	1	19	20
C	0	18	18
	1	37	38
3. (1.5 mo)			
I	0	19	19
C	2	16	18
	2	35	37
4. (3 mo)			
I	0	17	17
C	1	14	15
	1	31	32
5. (4.5 mo)			
I	1	15	16
C	0	12	12
	1	27	28
6. (4.8 mo)			
I	0	15	15
C	1	11	12
	1	26	27
7. (6.2 mo)			
I	0	14	14
C	1	10	11
	1	24	25
8. (10.5 mo)			
I	0	13	13
C	1	7	8
	1	20	21

[a]Number in parenthesis indicates time, t_j, of a death in either group

[b]Number of participant who died at time t_j

[c]Number of participants who are alive between time t_j and time t_{j+1}

[d]Number of participants who were at risk before death at time t_j ($R=D+A$)

where the vertical bars denote the absolute value. For example, applying the continuity correction reduces the MH statistic from 3.76 to 2.59.

Gehan [18] developed another procedure for comparing the survival experience of two groups of participants by generalizing the Wilcoxon rank statistic. The Gehan

statistic is based on the ranks of the observed survival times. The null hypothesis, $P_I(t) = P_C(t)$, is tested. The procedure, as originally developed, involved a complicated calculation to obtain the variance of the test statistic. Mantel [24] proposed a simpler version of the variance calculation, which is most often used.

The N_I observations from the intervention group and the N_C observations from the control group must be combined into a sequence of $N_C + N_I$ observations and ranked in ascending order. Each observation is compared to the remaining $N_C + N_I - 1$ observation and given a score U_i which is defined as follows:

U_i = (number of observations ranked definitely less than the ith observation) − (number of observations ranked definitely greater than the ith observation).

The survival outcome for the ith participant will certainly be larger than that for participants who died earlier. For censored participants, it cannot be determined whether survival time would have been less or greater than the ith observation. This is true whether the ith observation is a death or a loss. Thus, the first part of the score U_i assesses how many deaths definitely preceded the ith observation. The second part of the U_i score considers whether the current, ith, observation is a death or a loss. If it is a death, it definitely precedes all later ranked observations regardless of whether the observations correspond to a death or a loss. If the ith observation is a loss, it cannot be determined whether the actual survival time will be less than or greater than any succeeding ranked observation, since there was no opportunity to observe the ith participant completely.

Table 15.5 ranks the 40 combined observations ($N_C = 20$, $N_I = 20$) from the example used in the discussion of the Mantel–Haenszel statistic. The last 19 observations were all censored at 12 months of follow-up, 7 in the control group, and 12 in the intervention group. The score U_1 is equal to the zero observations that were definitely less than 0.5 months, minus the 39 observations that were definitely greater than 0.5 months, or $U_1 = -39$. The score U_2 is equal to the one observation definitely less than the loss at 0.6 months, minus none of the observations that will be definitely greater, since at 0.6 months the observation was a loss, or $U_2 = 1$. U_3 is equal to the one observation (0.5 months) definitely less than 1.0 month minus the 37 observations definitely greater than 1.0 month giving $U_3 = 36$. The last 19 observations will have scores of 9 reflecting the nine deaths which definitely precede censored observations at 12.0 months.

The Gehan statistic, G, involves the scores U_i and is defined as

$$G = W^2/V(W)$$

where $W = \Sigma U_i$, for (U_i's in control group only) and

$$V(W) = \frac{N_C N_I}{(N_C + N_I)(N_C + N_I - 1)} \sum_{i=1}^{N_C + N_I} (U_i^2)$$

The G statistic has approximately a chi-square distribution with one degree of freedom [18, 24]. Therefore, the critical value is 3.84 at the 5% significance level and 6.63 at the 1% level. In the example, $W = -87$ and the variance $V(W) = 2{,}314.35$.

Table 15.5 Example of Gehan statistics scores U_i for intervention (I) and control (C) groups

Observation	Ranked observed time	Group	Definitely less	Definitely more	U_i
1	0.5	C	0	39	−39
2	(0.6)[a]	C	1	0	1
3	1.0	I	1	37	−36
4	1.5	C	2	35	−33
5	1.5	C	2	35	−33
6	(1.6)	I	4	0	4
7	(2.0)	C	4	0	4
8	(2.4)	I	4	0	4
9	3.0	C	4	31	−27
10	(3.5)	C	5	0	5
11	(4.0)	C	5	0	5
12	(4.2)	I	5	0	5
13	4.5	I	5	27	−22
14	4.8	C	6	26	−20
15	(5.8)	I	7	0	7
16	6.2	C	7	24	−17
17	(7.0)	I	8	0	8
18	(8.5)	C	8	0	8
19	(9.0)	C	8	0	8
20	10.5	C	8	20	−12
21	(11.0)	I	9	0	9
22–40	(12.0)	12I, 7C	9	0	9

[a]Parentheses indicate censored observations

Thus, $G=(-87)^2/2{,}314.35=(87)^2/2{,}314.35$ or 3.27 for which the p-value is equal to 0.071. This is compared with the p value of 0.052 obtained using the Mantel–Haenszel statistic.

The Gehan statistic assumes the censoring pattern to be equal in the two groups. Breslow [25] considered the case in which censoring patterns are not equal and used the same statistic G with a modified variance. This modified version should be used if the censoring patterns are radically different in the two groups. Peto and Peto [26] also proposed a version of a censored Wilcoxon test. The concepts are similar to what has been described for Gehan's approach. However, most software packages now use the Breslow or Peto and Peto versions.

Generalizations

The general methodology of comparing two survival curves using this methodology has been further evaluated [27–32]. These two tests by Mantel–Haenzel and Gehan, can be viewed as a weighted sum of the difference between observed number of

events and the expected number at each unique event time [7, 27]. Consider the previous equation for the logrank test and rewrite the numerator as

$$W = \sum_{j=1}^{K} w_j [a_j - E(a_j)]$$

where

$$V(W) = \sum_{j=1}^{K} w_j^2 \frac{(a_j + c_j)(b_j + d_j)(a_j + b_j)(c_j + d_j)}{n_j^2 (n_j - 1)}$$

and w_j is a weighting factor. The test statistics $W^2/V(W)$ has approximately a chi-square distribution with one degree of freedom or equivalently $W\sqrt{V(W)}$ has approximately a standard normal distribution. If $w_i = 1$, we obtain the Mantel–Haenszel or logrank test. If $w_i = n_j/(N+1)$, where $N = N_C + N_I$ or the combined sample size, we obtain the Gehan version of the Wilcoxon test. Tarone and Ware [27] pointed out that the Mantel–Haenszel and Gehan are only two possible statistical tests. They suggested a general weight function $w_i = [n_j/(N+1)]^\theta$ where $0 \le \theta \le 1$. In particular, they suggested that $\theta = 0.5$. Prentice [29] suggested a weight $w_j = \prod_{i=1}^{j} n_i / (n_i + d_i)$ where $d_i = (a_i + c_i)$ which is related to the product limit estimator at t_j as suggested by Peto and Peto [26]. Harrington and Fleming [32] generalize this further by suggesting weights $w_j = \left\{\prod_{i=1}^{j} n_i / (n_i + d_i)\right\}^\rho$ for $\rho \ge 0$.

All of these methods give different weights to the various parts of the survival curve. The Mantel–Haenszel or logrank statistic is more powerful for survival distributions of the exponential form where $\lambda_I(t) = \theta\lambda_C(t)$ or $S_I(t) = \{S_C(t)\}^\theta$ where $\theta \ne 1$ [24]. The Gehan type statistic [18], on the other hand, is more powerful for survival distributions of the logistic form $S(t,\theta) = e^{t+\theta}/(1 + e^{t+\theta})$. In actual practice, however, the distribution of the survival curve of the study population is not known. When the null hypothesis is not true, the Gehan type statistic gives more weight to the early survival experience, whereas the Mantel–Haenszel weights the later experience more. Tarone and Ware [27] indicate that other possible weighting schemes, which are intermediate to these two statistics, could be proposed [27, 32]. Thus, when survival analysis is done, it is certainly possible to obtain different results using different weighting schemes depending on where the survival curves separate, if they indeed do so. The logrank test is the standard in many fields such as cancer and heart disease. The condition $\lambda_I(t) = \theta\lambda_C(t)$ says that risk of the event being studied in the intervention is a constant multiple of the hazard $\lambda_C(t)$. That is, the hazard rate in one arm is proportional to the other and so the logrank test is best for testing proportional hazards. This idea is appealing and is approximately true for many studies.

There has been considerable interest in asymptotic (large sample) properties of rank tests [28, 30] as well as comparisons of the various analytic methods [31]. While there exists an enormous literature on survival analysis, the basic concepts of rank tests can still be appreciated by the methods described above.

Earlier, we discussed using an exponential model to summarize a survival curve where the hazard rate λ determines the survival curve. If we can assume that the hazard rate is reasonably constant during the period of follow-up for the intervention and the control group, then comparison of hazard rates is a comparison of survival curves [4]. The most commonly used comparison is the ratio of the hazards, $R = \lambda_I / \lambda_C$. If the ratio is unity, the survival curves are identical. If R is greater than one, the intervention hazard is greater than control so the intervention survival curve falls below the standard curve. That is, the intervention is worse. On the other hand, if R is less than one, the control group hazard is larger, the control group survival curve falls below the intervention curve, and intervention is better.

We can estimate the hazard ratio by comparing the ratio of total observed events (O) divided by expected number of events (E) in each group; that is, the estimate of R can be expressed as

$$\hat{R} = \frac{O_I / E_I}{O_C / E_C}$$

That is, $O_I = \sum a_i$, $O_C = \sum (b_i)$, $E_I = \sum E(a_i)$, and $E_C = \sum E(b_i)$. Confidence intervals for the odds ratio R are most easily determined by constructing confidence intervals for the log of the odds ratio $\ln R$ [33]. The 95% confidence interval for $\ln R$ is $K - 1.96 / \sqrt{V}$ to $K + 1.96 / \sqrt{V}$ where $K = (O_I - E_I)/V$ and V is the variance as defined in the logrank or Mantel–Haenszel statistics. (That is, V equals $V(a_i)$.) We then connect confidence intervals for $\ln R$ to confidence intervals for R by taking antilogs of the upper and lower limit. If the confidence interval excludes unity, we could claim superiority of either intervention or control depending on the direction. Hazard ratios not included in the interval can be excluded as likely outcome summaries of the intervention. If the survival curves have relatively constant hazard rates, this method provides a nice summary and complements the Kaplan–Meier estimates of the survival curves.

Covariate Adjusted Analysis

Previous chapters have discussed the rationale for taking stratification into account. If differences in important covariates or prognostic variables exist at entry between the intervention and control groups, an investigator might be concerned that the analysis of the survival experience is influenced by that difference. In order to adjust for these differences in prognostic variables, she could conduct a stratified analysis or a covariance type of survival analysis. If these differences are not important in the analysis, the adjusted analysis will give approximately the same results as the unadjusted.

Three basic techniques for stratified survival analysis are of interest. The first compares the survival experience between the study groups within each stratum, using the methods described in the previous section. By comparing the results from

each stratum, the investigator can get some indication of the consistency of results across strata and the possible interaction between strata and intervention.

The second and third methods are basically adaptations of the Mantel–Haenszel and Gehan statistics, respectively, and allow the results to be accumulated over the strata. The Mantel–Haenszel stratified analysis involves dividing the population into S strata and within each stratum j, forming a series of 2×2 tables for each K_j event, where K_j is the number of events in stratum j. The table for the ith event in the jth stratum would be as follows:

	Event	Alive	
Intervention	a_{ij}	b_{ij}	$a_{ij}+b_{ij}$
Control	c_{ij}	d_{ij}	$c_{ij}+d_{ij}$
	$a_{ij}+c_{ij}$	$b_{ij}+d_{ij}$	n_{ij}

The entries a_{ij}, b_{ij}, c_{ij}, and d_{ij} are defined as before and

$$E(a_{ij}) = (a_{ij} + c_{ij})(a_{ij} + b_{ij}) / n_{ij}$$

$$V(a_{ij}) = \frac{(a_{ij} + c_{ij})(b_{ij} + d_{ij})(a_{ij} + b_{ij})(c_{ij} + d_{ij})}{n_{ij}^2(n_{ij} - 1)}$$

Similar to the non-stratified case, the Mantel–Haenszel statistic is

$$\mathrm{MH} = \left\{\sum_{j=1}^{S}\sum_{i=1}^{K_j} a_{ij} - E(a_{ij})\right\}^2 / \sum_{j=1}^{S}\sum_{i=1}^{K_j} V(a_{ij})$$

which has a chi-square distribution with one degree of freedom. Analogous to the Mantel–Haenszel statistic for stratified analysis, one could compute a Gehan statistic W_j and $V(W_j)$ within each stratum. Then an overall stratified Gehan statistic is computed as

$$G = \left\{\sum_{j=1}^{S} W_j\right\}^2 / \sum_{j=1}^{S} V(W_j)$$

which also has chi-square statistic with one degree of freedom.

If there are many covariates, each with several levels, the number of strata can quickly become large, with few participants in each. Moreover, if a covariate is continuous, it must be divided into intervals and each interval assigned a score or rank before it can be used in a stratified analysis. Cox [34] proposed a regression model which allows for analysis of censored survival data adjusting for continuous as well as discrete covariates, thus avoiding these two problems.

One way to understand the Cox regression model is to again consider a simpler parametric model. If one expresses the probability of survival to time t, denoted

$S(t)$, as an exponential model, then $S(t) = e^{-\lambda t}$ where the parameter, λ, is called the force of mortality or the hazard rate as described earlier. The larger the value of λ, the faster the survival curve decreases. Some models allow the hazard rate to change with time, that is $\lambda = \lambda(t)$. Models have been proposed [35–37] which attempt to incorporate the hazard rate as a linear function of several baseline covariates, $x_1, x_2, \ldots, x_p$ that is, $\lambda(x_1, x_2, \ldots, x_p) = b_1x_1 + b_2x_2 + \cdots + b_px_p$. One of the covariates, say x_1, might represent the intervention and the others, for example, might represent age, sex, performance status, or prior medical history. The coefficient, b_1, then would indicate whether intervention is a significant prognostic factor, i.e., remains effective after adjustment for the other factors. Cox [34] suggested that the hazard rate could be modeled as a function of both time and covariates, denoted $\lambda(t, x_1, x_2, \ldots, x_p)$. Moreover, this hazard rate could be represented as the product of two terms, the first representing an unadjusted force of mortality $\lambda_0(t)$ and the second the adjustment for the linear combination of a particular covariate profile. More specifically, the Cox proportional hazard model assumes that

$$\lambda(t, x_1, x_2, \ldots, x_p) = \lambda_0(t) \exp(b_1x_1 + b_2x_2 + \cdots + b_px_p)$$

That is, the hazard $\lambda(t, x_1, x_2, \ldots, x_n)$ is proportional to an underlying hazard function $\lambda_0(t)$ by the specific factor $\exp(b_1x_1 + b_2x_2 \ldots)$. From this model, we can estimate an underlying survival curve $S_0(t)$ as a function of $\lambda_0(t)$. The survival curve for participants with a particular set of covariates X, $S(t,x)$ can be obtained as $S(t, x) = [S_0(t)]^{\exp(b_1x_1 + b_2x_2 + \cdots)}$. Other summary test statistics from this model are also used. The estimation of the regression coefficients $b_1, b_2, \ldots, b_p$ is complex, requiring non-linear regression methods, and goes beyond the scope of this text. Many elementary texts on biostatistics [3, 4, 6] or review articles [7] present further details. A more advanced discussion may be found in Kalbfleish and Prentice [8] or Fleming and Harrington [11]. Programs exist in many statistical computing packages which provide these estimates and summary statistics to evaluate survival curve comparisons. Despite the complexity of the parameter estimation, this method is widely applied and has been studied extensively [38–47]. Pocock et al. [47] demonstrate the value of some of these methods with cancer data. For the special case where group assignment is the only covariate, the Cox model is essentially equivalent to the Mantel–Haenszel statistic.

One issue that is sometimes raised is whether the hazard rates are proportional over time. Methods such as the Mantel–Haenszel logrank test or the Cox proportional hazards model are the most efficient or powerful when the hazards are proportional [11]. However, these methods are still valid as long as the survival curves for example do not cross. In that case, which intervention is better depends on what time point is being referenced. With that exception, these methods are valid without the proportional hazards assumption, although perhaps not as powerful as when the hazards are proportional. That is, if a significant difference is found between two survival curves when the hazards are not proportional, the two curves are still significantly different. For example, time to event curves are shown in Chap. 17. Figure 17.2a shows three curves for comparison of two medical devices with best

medical or pharmacologic care. These three curves do not have proportional hazards but the comparisons are still valid and in fact the two devices demonstrate statistically significant superiority over the best medical care arm. The survival curves do not cross although are close together in the early months of follow-up.

The techniques described in this chapter as well as the extensions or generalizations referenced are powerful tools in the analysis of survival data. Perhaps none is exactly correct for any given set of data, but experience indicates they are fairly robust and quite useful.

References

1. Brown BW, Hollander M. *Statistics: A Biomedical Introduction*. New York: John Wiley and Sons, 1977.
2. Armitage P. *Statistical Methods in Medical Research*. New York: John Wiley and Sons, 1977.
3. Breslow N. Comparison of survival curves. In Buyse B, Staquet M, Sylvester R (eds). *The Practice of Clinical Trials in Cancer*. Oxford: Oxford University Press, 1982.
4. Altman DG. *Practical Statistics for Medical Research*. New York: Chapman and Hall, 1991, pp. 383–392.
5. Woolson R. *Statistical Methods for the Analysis of Biomedical Data*. New York: John Wiley and Sons, 1987.
6. Fisher L, VanBelle G. *Biostatistics: A Methodology for the Health Sciences*. New York: John Wiley and Sons, 1983.
7. Crowley J, Breslow N. Statistical analysis of survival data. *Annu Rev Public Health* 1984;5:385–411.
8. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. New York: John Wiley and Sons, 1980.
9. Miller RG, Jr. *Survival Analysis.* New York: John Wiley and Sons, 1981.
10. Cox DR, Oakes D. *The Analysis of Survival Data.* New York: Chapman and Hall, 1984.
11. Fleming T, Harrington D. *Counting Processes and Survival Analysis*. New York: John Wiley and Sons, 1991.
12. Kaplan E, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958;53:457–481.
13. Cutler S, Ederer F. Maximum utilization of the lifetable method in analyzing survival. *J Chronic Dis* 1958;8:699–712.
14. Greenwood M. The natural duration of cancer. *Rep Publ Health Med Subj* 1926;33:1–26.
15. Thomas DG, Breslow N, Gart J. Trend and homogeneity analysis of proportions and life table data. *Comput Biomed Res* 1977;10:373–381.
16. Nelson W. Hazard plotting for incomplete failure data. *J Qual Technol* 1969;1:27–52.
17. Brookmeyer R, Crowley J. A confidence interval for the median survival time. *Biometrics* 1982;38:29–42.
18. Gehan E. A generalized Wilcoxon test for comparing arbitrarily single censored samples. *Biometrika* 1965;52:203–223.
19. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* 1966;50:163–170.
20. Cochran W. Some methods for strengthening the common χ^2 tests. *Biometrics* 1954;10: 417–451.
21. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959;22:719–748.

22. Peto R, Pike MC. Conservatism in the approximation $\Sigma(0-E)^2/E$ in the logrank test for survival data or tumor incidence data. *Biometrics* 1973;29:579–584.
23. Crowley J, Breslow N. Remarks on the conservatism of $\Sigma(0-E)^2/E$ in survival data. *Biometrics* 1975;31:957–961.
24. Mantel N. Ranking procedures for arbitrarily restricted observations. *Biometrics* 1967;23:65–78.
25. Breslow N. A generalized Kruskal–Wallis test for comparing *K* samples subject to unequal patterns of censorship. *Biometrika* 1970;57:579–594.
26. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *J R Stat Soc Ser A* 1972;135:185–207.
27. Tarone R, Ware J. On distribution-free tests for equality of survival distributions. *Biometrika* 1977;64:156–160.
28. Oakes D. The asymptotic information in censored survival data. *Biometrika* 1977;64:441–448.
29. Prentice RL. Linear rank tests with right censored data. *Biometrika* 1978;65:167–179.
30. Schoenfeld D. The asymptotic properties of non-parametric tests for comparing survival distributions. *Biometrika* 1981;68:316–319.
31. Leurgans SL. Three classes of censored data rank tests: strengths and weaknesses under censoring. *Biometrika* 1983;70:651–658.
32. Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. *Biometrika* 1982;69:553–566.
33. Simon R. Confidence intervals for reporting results of clinical trials. *Ann Intern Med* 1986;105:429–435.
34. Cox DR. Regression models and life tables. *J R Stat Soc Series B Stat Methodol* 1972;34:187–202.
35. Zelen M. Application of exponential models to problems in cancer research. *J R Stat Soc Ser A* 1966;129:368–398.
36. Feigl P, Zelen M. Estimation of exponential survival probabilities with concomitant information. *Biometrics* 1965;21:826–838.
37. Prentice RL, Kalbfleisch JD. Hazard rate models with covariates. *Biometrics* 1979;35:25–39.
38. Kalbfleisch JD, Prentice RL. Marginal likelihoods based on Cox's regression and life model. *Biometrika* 1973;60:267–278.
39. Breslow N. Covariance analysis of censored survival data. *Biometrics* 1974;30:89–99.
40. Breslow N. Analysis of survival data under the proportional hazards model. *Int Stat Rev* 1975;43:45–58.
41. Kay R. Proportional hazard regression models and the analysis of censored survival data. *J R Stat Soc Ser C Appl Stat* 1977;26:227–237.
42. Prentice RL, Gloeckler LA. Regression analysis of grouped survival data with application to breast cancer. *Biometrics* 1978;34:57–67.
43. Efron B. The efficiency of Cox's likelihood function for censored data. *J Am Stat Assoc* 1977;72:557–565.
44. Tsiatis AA. A large sample study of Cox's regression model. *Ann Stat* 1981;9:93–108.
45. Schoenfeld D. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika* 1980;67:145–153.
46. Storer BE, Crowley J. Diagnostics for Cox regression and general conditional likelihoods. *J Am Stat Assoc* 1985;80:139–147.
47. Pocock SJ, Gore SM, Kerr GR. Long term survival analysis: the curability of breast cancer. *Stat Med* 1982;1:93–104.

Chapter 16
Monitoring Response Variables

The investigator's ethical responsibility to the study participants demands that results in terms of safety and clinical benefit be monitored during trials. If data partway through the trial indicate that the intervention is harmful to the participants, early termination of the trial should be considered. If these data demonstrate a clear benefit from the intervention, the trial may also be stopped early because to continue would be unethical to the participants in the control group. In addition, if differences in primary and possibly secondary response variables are so unimpressive that the prospect of a clear result is extremely unlikely, it may not be justifiable in terms of time, money, and effort to continue the trial. Also, monitoring of response variables can identify the need to collect additional data to clarify questions of benefit or toxicity that may arise during the trial. Finally, monitoring may reveal logistical problems or issues involving data quality that need to be promptly addressed. Thus, there are ethical, scientific, and economic reasons for interim evaluation of a trial [1–3]. In order to fulfill the monitoring function, the data must be collected and processed in a timely fashion as the trial progresses. Data monitoring would be of limited value if conducted only at a time when all or most of the data had been collected. The specific issues related to monitoring of recruitment, adherence, and quality control are covered in other chapters and will not be discussed here. The data monitoring committee process has been described in detail [4] as have case studies representing trials, which were terminated for benefit, harm, or futility [5]. One of the earliest discussions of the basic rationale for data monitoring was included in a report of a committee initiated at the request of the council advisory to the then National Heart Institute and chaired by Bernard Greenberg [1]. This report outlined a clinical trial model depicted in Fig. 16.1, variations of which have been implemented widely by institutes at the National Institutes of Health (NIH). The key components are the Steering Committee, the Statistical and Data Coordinating Center, the Clinics, and the Data Monitoring Committee. Later the pharmaceutical and device industries [6] adopted a modified version of this NIH model, depicted in Fig. 16.2. The main modification was to separate the Statistical Data Coordinating Center into a Statistical Data Analysis Center and a Data Coordinating Center. Many of the early

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_16, © Springer Science+Business Media, LLC 2010

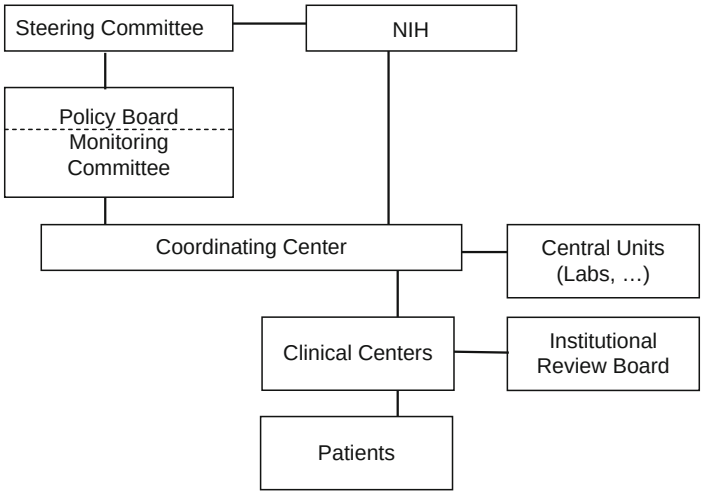


Fig. 16.1 The NIH clinical trial model [6]

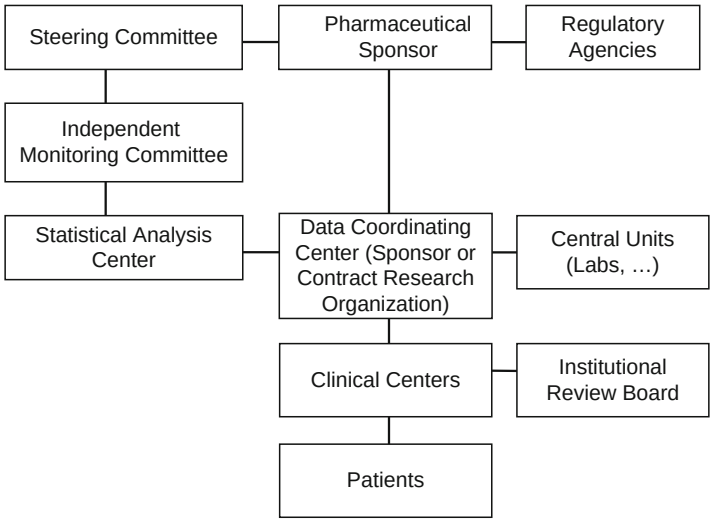


Fig. 16.2 The industry-modified clinical trial model [6]

experiences have been described and formed the basis of current practice [7–34], particularly in trials of cardiovascular disease [35–37].

In 2000, the death of a young person undergoing gene transfer as part of a research protocol brought a great deal of attention to the process of monitoring trials and the reporting requirements. As a result, the U.S. Secretary of Health and Human Resources issued a requirement that all trials sponsored by the NIH or under the regulatory review of the Food and Drug Administration (FDA) have a monitoring plan [38–40].

For some trials this entails an independent monitoring committee. The NIH developed policies regarding trial monitoring that required almost all Phase III trials to have a monitoring committee as outlined in their guidelines. The FDA guidelines finalized in 2006, recommend an independent monitoring committee for trials involving high risk patients or with novel or potentially high risk interventions.

A survey of monitoring practices conducted by the DAMOCLES group found that the roles of monitoring committees varied widely across trials, sponsors, and regions. While there was a general agreement about the types of trials that needed formal monitoring committees, there was not a uniform practice or policy as to their function [41]. The principles and fundamentals expressed in this book reflect the experience of the authors in monitoring numerous trials since the early 1970s.

Fundamental Point

During the trial, response variables need to be monitored for early dramatic benefits or potential harmful effects. Preferably, monitoring should be done by a person or group independent of the investigator. Although many techniques are available to assist in monitoring, none of them should be used as the sole basis in the decision to stop or continue the trial.

Monitoring Committee

Keeping in mind the scientific, ethical, and economic rationales, data and safety monitoring is not simply a matter of looking at tables or results of statistical analysis of the primary outcome. Rather, it is an active process in which additional tabulations and analysis are suggested and evolve as a result of ongoing review. Monitoring also involves an interaction between the individuals responsible for collating, tabulating, and analyzing the data. For single center studies, the monitoring responsibility could, in principle, be assumed by the investigator. However, he may find himself in a difficult situation. While monitoring the data, he may discover that the results trend in one direction or the other while participants are still being enrolled and/or treated. Presumably, he recruits participants to enter a trial on the basis that he favors neither intervention nor control, a state of clinical equipoise [42]. Knowing that a trend exists may make it difficult for him to continue enrolling participants. It is also difficult for the investigator to follow, evaluate, and care for the participants in an unbiased manner knowing that a trend exists. Furthermore, the credibility of the trial is enhanced if, instead of the investigator, an independent person monitors the response variable data. Because of these considerations, we recommend that the individuals who monitor a clinical trial have no formal involvement with the participants or the investigators, although some disagree [26–28].

Except for small, short-term studies, when one or two knowledgeable individuals may suffice, the responsibility for monitoring response variable data is usually placed with an independent group with expertise in various disciplines [4–6]. The independence protects the members of the monitoring committee from being influenced in the decision-making process by investigators, participants as well as federal or industry sponsors. The committee would usually include experts in the relevant clinical fields or specialties, individuals with experience in the conduct of clinical trials, epidemiologists, biostatisticians knowledgeable in design and analysis, and often a bioethicist and participant advocate. While we will describe statistical procedures that are often helpful in evaluating interim results, the decision process to continue, terminate a trial early, or modify the design is invariably complex and no single statistical procedure is adequate to address all these complexities. Furthermore, no single individual is likely to have all the experiences and expertises to deal with these issues. Thus, as was recommended in the Greenberg Report [1], we suggest that the independent monitoring committee have a multidisciplinary membership.

The first priority of the monitoring committee must be to ensure the safety of the participants in the trial. The second priority is to the investigators and the Institutional Review Boards or ethics committees, who place an enormous trust in the monitoring committee both to protect their participants from harm and to ensure the integrity of the trials. Third, the monitoring committee has a responsibility to the sponsor of the trial, whether it be federal or private. Finally, the monitoring committee provides a service to the drug or device regulatory agency, especially for trials which are utilizing drugs, biologics or devices which still have investigational status.

Although many formats for monitoring committee meetings have been used, one that we recommend allows for exchange of information by all relevant parties and for appropriate confidential and independent review [4, 33]. The format utilizes an open session, a closed session, and an executive session. The open session enables interaction between investigator representatives such as the study principal investigator or chair, the sponsor, the statistical center, the relevant industrial participants, and the monitoring committee. It is uncommon for a regulatory agency to participate in a meeting. In this session, issues of participant recruitment, data quality, general adherence, toxicity issues, and any other logistical matter that may affect either the conduct or outcome of the trial are considered. After a thorough discussion, the monitoring committee would go into a closed session where analyses of the confidential blinded outcome data are reviewed. This review would include comparison by intervention groups of baseline variables, primary or secondary variables, safety or adverse outcome variables, adherence measures for the entire group, and examinations of any relevant subgroups. Following this review, the monitoring committee would move into an executive session where decisions about continuation, termination or protocol modification are made. These different sessions may be formally or informally divided, depending on who attends the monitoring committee meetings. Regardless of how formal, most monitoring committee meetings have such

components. This particular model, for example, has been used extensively in NIH-sponsored AIDS trials [33].

Before a trial begins and the first monitoring committee meeting is scheduled, it must be decided specifically who attends the various sessions, as outlined above. In general, attendance should be limited to those who are essential for proper monitoring. As noted, it is common for the study principal investigator and sponsor representatives to attend the open session. If the principal investigator does not care for participants in the trial, that individual will sometimes attend the closed session, although there is variation in that practice. If the study is sponsored by industry, independence and credibility of the study are best served by no industry attendance at the closed session. Industry sponsored trials that are also managed and analyzed by industry will require a biostatistician from the sponsor who prepares the monitoring report to attend. The company statistician must have a "firewall" separating her from other colleagues at the company, something that may be difficult to achieve or be convincing to others. However, a common practice for industry-sponsored pivotal Phase III trials is for a separate statistical analysis center to provide the interim analyses and report to the independent monitoring committee [6]. This practice reduces the possibility or perception that interim results are known within the industry sponsor, or the investigator group. Regulatory agency representatives usually do not attend the closed session because being involved in the monitoring decision may affect their regulatory role, should the product be submitted for subsequent approval.

An executive session should involve only the voting members of the monitoring committee, although the independent statistician who provided the data report may also attend. There are many variations of this general outline, including a merger of the closed and executive session since attendance may involve the same individuals.

Most monitoring committees evaluate one, or perhaps two, clinical trials. When a trial is completed, that monitoring committee is dissolved. However, as exemplified by cancer and AIDS, ongoing networks of clinical centers conduct many trials concurrently [12, 14, 26–28, 33]. Cancer trial cooperative groups may conduct trials across several cancer sites, such as breast, colon, lung or head, and neck at any given time, and even multiple trials for a given site depending upon the stage of the cancer or other risk factors. The AIDS trial networks in the United States have likewise conducted trials simultaneously in AIDS patients at different stages of the disease. In these areas, monitoring committees may follow the progress of several trials. In such instances, a much disciplined agenda and a standardized format of the data report enhance the efficiency of the review. Regardless of the model, the goals and procedures are similar.

Another factor that needs to be resolved before the start of a trial is how the intervention or treatment comparisons will be presented to the monitoring committee. In some trials, the monitoring committee knows the identity of the interventions in each table or figure of the report. In other trials, the monitoring committee is blinded throughout the interim monitoring. In order to achieve this, data reports

have complex labeling schemes, such as A versus B for baseline tables, C versus D for primary outcomes, E versus F for toxicity, and G versus H for various laboratory results. While this degree of blinding may enhance objectivity, it may conflict with the monitoring committee's primary purpose of protecting the participants in the trial from harm or unnecessary continuation. To assess the progress of the trial, the harm and benefit profile of the intervention must be well understood and the possible tradeoffs weighed. If each group of tables is labeled by a different code, the committee cannot easily assess the overall harm/benefit profile of the intervention, and thus may put participants at unnecessary risk or continue a trial beyond the point at which benefit outweighs risks. Such complex coding schemes also increase the chance for errors in labeling. A reasonable compromise is to label all tables consistently, such as arm A and B, or at most by two codes, with the understanding that the committee can become unblinded. Thus, if there are no trends in either benefit or harm, which is likely to be the case early in a trial, there is no overwhelming reason to know the identity of groups A and B. When trends begin to emerge in either direction, the monitoring committee should have full knowledge of the group identities [43].

No simple formula can be given for how often a monitoring committee should meet. The frequency may vary depending on the phase of the trial [3–5, 44]. Participant recruitment, follow-up, and closeout phases require different levels of activity. Meetings should not be so frequent that little new data are accumulated in the interim, given the time and expense of convening a committee. If potential toxicity of one of the interventions becomes an issue during the trial, special meetings may be needed. In many long-term clinical trials, the monitoring committees have met regularly at 4- to 6-month intervals, with additional meetings or telephone conferences as needed. In some circumstances, an annual review may be sufficient. However, less frequent review is not recommended since too much time may elapse before a serious adverse effect is uncovered. As described later, another strategy is to schedule monitoring committee meetings when approximately 10, 25, 50, 75, and 100% of the primary outcomes have been observed, or some similar pattern. Thus, there might be an early analysis to check for serious immediate adverse effects with later analyses to evaluate evidence of intervention benefit or harm. Other approaches provide for additional in-between analyses if strong, but as yet non-significant trends emerge. Between committee meetings, the person or persons responsible for collating, tabulating, and analyzing the data assume the responsibility for monitoring unusual situations which may need to be brought to the attention of the monitoring committee.

A monitoring committee often reviews the data for the last time before the data file is closed, and may never see the complete data analysis except as it appears in the publication. There is currently no consistent practice as to whether a monitoring committee meets to review the final complete data set. From one perspective, the trial is over and there is no need for the committee to meet since early termination or protocol modification is no longer an option. From another perspective, the committee has become very familiar with the data, including issues of potential concern, and thus may have insight to share with the investigators and study sponsors.

Some trials have scheduled this final meeting to allow the monitoring committee to see the final results before they are presented at a scientific meeting or published. Based on our experience, we strongly recommend this latter approach. There is a great deal to be gained for the trial and the investigators at a very modest cost.

Other remaining issues still need to be resolved. For example, if a worrisome safety trend or a significant finding is not reported clearly or at all in the primary publication, what are the scientific, ethical, and legal obligations for the monitoring committee to comment on what is not reported? Suppose the committee differs substantially in the interpretation of the primary or safety outcomes? What is the process for resolving differences between it and the investigators or sponsor? These are important questions and the answers are not simple or straightforward, yet are relevant for science and ethics.

Repeated Testing for Significance

In the discussion on sample size (Chap. 8) the issue of testing several hypotheses was raised and referred to as the "multiple testing" problem. Similarly, repeated significance testing of accumulating data is essential to the monitoring function has statistical implications [45–51]. If the null hypothesis, H_0, of no difference between two groups is, in fact, true, and repeated tests of that hypothesis are made at the same level of significance using accumulating data, the probability that, at some time, the test will be called significant by chance alone is larger than the significance level selected. That is, the rate of incorrectly rejecting the null hypothesis, or making a false positive error, will be larger than what is normally considered acceptable. Trends may emerge and disappear, especially early in the trial, and caution must be used. Here, we present the issue from a classical frequentist view point although other statistical approaches, such as the Bayesian methods that are discussed briefly near the end of this chapter, exist.

In a clinical trial in which the participant response is known relatively soon after entry, the difference in rates between two groups may be compared repeatedly as more participants are added and the trial continues. The usual test statistic for comparing two proportions used is the chi-square test or the equivalent normal test statistic. The null hypothesis is that the true response rates or proportions are equal. If a significance level of 5% is selected and the null hypothesis, H_0, is tested only once, the probability of rejecting H_0 if it is true is 5% by definition. However, if H_0 is tested twice, first when one-half of the data are known and then when all the data are available, the probability of incorrectly rejecting H_0 is increased from 5 to 8% [49]. If the hypothesis is tested five times, with one-fifth of the participants added between tests, the probability of finding a significant result if the usual statistic for the 5% significance level is used becomes 14%. For ten tests, this probability is almost 20%.

In a clinical trial in which long-term survival experience is the primary outcome, repeated tests might be done as more information becomes known about the enrolled

participants. Canner [24] performed computer simulations of such a clinical trial in which both the control group and intervention group event rates were assumed to be 30% at the end of the study. He performed 2,000 replications of this simulated experiment. He found that if 20 tests of significance are done within a trial, the chance of crossing the 5% significance level boundaries (i.e., $Z=\pm 1.96$) is, on the average, 35%. Thus, whether one calculates a test statistic for comparing proportions or for comparing time to event data, repeated testing of accumulating data without taking into account the number of tests increases the overall probability of incorrectly rejecting H_0. If the repeated testing continues indefinitely, the null hypothesis is certain to be rejected eventually. Although it is unlikely that a large number of repeated tests will be done, even five or ten can lead to a misinterpretation of the results of a trial if the multiple testing issues are ignored.

A classic illustration of the repeated testing problem is provided by the Coronary Drug Project (CDP) for the clofibrate versus placebo mortality comparison, shown in Fig. 16.3 [24, 51]. This figure presents the standardized mortality comparisons over the follow-up or calendar time of the trial. The two horizontal lines indicate the conventional value of the test statistic, corresponding to a two-sided 0.05 significance level, used to judge statistical significance for studies where the comparison is made just one time. It is evident that the trends in this comparison emerge and weaken throughout, coming close or exceeding the conventional critical values on five monitoring occasions. However, as shown in Fig. 16.4, the mortality curves at the end of the trial are nearly identical, corresponding to the very small standardized statistic at the end of the Fig. 16.3. The monitoring committee for this trial took into consideration the repeated testing problem and did not terminate this trial early because the conventional values were exceeded.

For ethical, scientific, and economic reasons, all trials must be monitored so as not to expose participants to possible harm unnecessarily, waste precious fiscal and

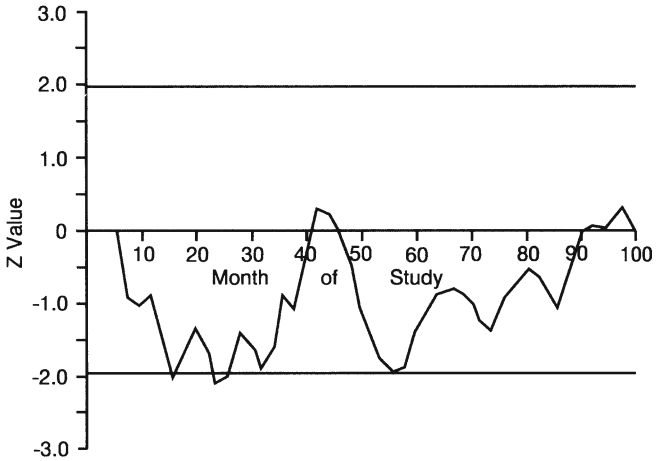


Fig. 16.3 Interim survival analyses comparing mortality in clofibrate- and placebo-treated participants in the Coronary Drug Project. A positive Z value favors placebo [7]

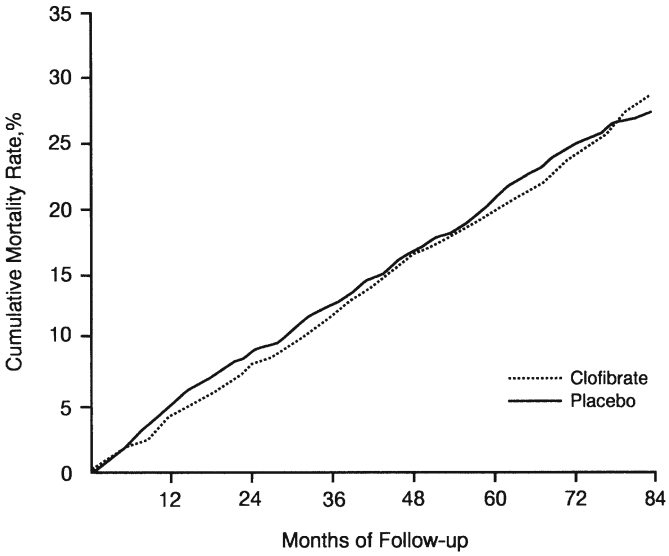


Fig. 16.4 Cumulative mortality curves comparing clofibrate- and placebo-treated participants in the Coronary Drug Project [7]

human resources, or miss opportunities to correct flaws in the design [1, 3–5]. However, in the process of evaluating interim results to meet these responsibilities, incorrect conclusions can be drawn by overreacting to emerging or non-emerging trends in primary, secondary or adverse effect outcomes. In general, the solution to multiple testing is to adjust the critical value used in each analysis so that the overall significance level for the trial remains at the desired level. It has been suggested that a trial should not be terminated early unless the difference between groups is very significant [3–5, 52]. More formal monitoring techniques are reviewed later in this chapter. They include the group sequential methods and stochastic curtailed sampling procedures.

Decision for Early Termination

There are five major valid reasons for terminating a trial earlier than scheduled [3–5, 7, 24]. First, the trial may show serious adverse effects in the entire intervention group or in a dominating subgroup. Second, the trial may indicate greater than expected beneficial effects. Third, it may become clear that a statistically significant difference by the end of the study is improbable. Fourth, logistical or data quality problem may be so severe that correction is not feasible or participant recruitment is far behind and not likely to achieve the target. Fifth, the question posed may have already been answered elsewhere or may no longer be sufficiently important. A few trials have been terminated because the sponsor decided the trial

was no longer a priority but this causes serious ethical dilemmas for investigators and leaves participants having contributed without getting an answer to the posed question.

For a variety of reasons, a decision to terminate a study early must be made with a great deal of caution and in the context of all pertinent data. A number of issues or factors must be considered thoroughly as part of the decision process:

1. Possible differences in prognostic factors between the two groups at baseline.
2. Any chance of bias in the assessment of response variables, especially if the trial is not double-blind.
3. The possible impact of missing data. For example, could the conclusions be reversed if the experience of participants with missing data from one group were different from the experience of participants with missing data from the other group?
4. Differential concomitant intervention and levels of participant adherence.
5. Potential adverse events and outcomes of secondary response variables in addition to the outcome of the primary response variable.
6. Internal consistency. Are the results consistent across subgroups and the various primary and secondary outcome measures? In a multicenter trial, the monitoring committee should assess whether the results are consistent across centers. Before stopping, the committee should make certain that the outcome is not due to unusual experience in only one or two centers.
7. In long-term trials, the experience of the study groups over time. Survival analysis techniques (Chap. 15) partly address this issue.
8. The outcomes of similar trials.
9. The impact of early termination on the credibility of the results and acceptability by the clinical community.

Some trials request the chair of the monitoring committee to review frequently serious adverse events, by intervention, to protect the safety of the participants. While such frequent informal, or even formal, review of the data is also subject to the problems of repeated testing or analyses, the adjustment methods presented are typically not applied. Also, safety may be measured by many response variables. Rather than relying on a single outcome showing a worrisome trend, a profile of safety measures might be required. Thus, the decision to stop a trial for safety reasons can be quite complex.

The early termination of a clinical trial can be difficult [3, 7, 8, 24, 52–57], not only because the issues involved may be complex and the study complicated but also because the final decision often lies with the consensus of a committee. The statistical methods discussed in this chapter are useful guides in this process but should not be viewed as absolute rules. A compilation of diverse monitoring experiences is available [5]. A few examples are described here to illustrate key points.

One of the earlier clinical trials conducted in the United States illustrates how controversial the decision for early termination may be. The University Group Diabetes Program (UGDP) was a placebo-control, randomized, double-blind trial designed to test the effectiveness of four interventions used in the treatment of

diabetes [58–61]. The primary measure of efficacy was the degree of retinal damage. The four interventions were: a fixed dose of insulin, a variable dose of insulin, tolbutamide and phenformin. After the trial was underway, study leaders formed a committee to review accumulating safety data. This committee membership consisted of individuals involved in the UGDP and external consultants. The tolbutamide group was stopped early because the monitoring committee thought the drug could be harmful and did not appear to have any benefit [58]. An excess in cardiovascular mortality was observed in the tolbutamide group as compared to the placebo group (12.7% vs. 4.9%) and the total mortality was in the same direction (14.7% vs. 10.2%). Analysis of the distribution of the baseline factors known to be associated with cardiovascular mortality revealed an imbalance, with participants in the tolbutamide group being at higher risk. This, plus questions about the classification of cause of death, drew considerable criticism. Later, the phenformin group was also stopped because of excess mortality in the control group (15.2% vs. 9.4%) [60]. The controversy led to a review of the data by an independent group of statisticians. Although they basically concurred with the decisions made by the UGDP monitoring committee [60], the debate over the study and its conclusion continued [61].

The decision-making process during the course of the CDP [62] a long-term randomized, double-blind, multicenter study that compared the effect on total mortality of several lipid-lowering drugs (high- and low-dose estrogen, dextrothyroxine, clofibrate, nicotinic acid) against placebo has been reviewed [5, 7, 51, 62, 64]. Three of the interventions were terminated early because of potential side effects and no apparent benefit. One of the issues in the discontinuation of the high dose estrogen and dextrothyroxine interventions [62, 63] concerned subgroups of participants. In some subgroups, the interventions appeared to cause increased mortality, in addition to having a number of other adverse events. In others, the adverse events were present, but mortality was only slightly reduced or unchanged. The adverse events were thought to more than outweigh the minimal benefit in selected subgroups. Also, positive subgroup trends in the dextrothyroxine arm were not maintained over time. After considerable debate, both interventions were discontinued. The low dose estrogen intervention [64] was discontinued because concerns over major toxicity. Furthermore, it was extremely improbable that a significant difference in a favorable direction for the primary outcome (mortality) could have been obtained had the study continued to its scheduled termination. Using the data available at the time, the number of future deaths in the control group was projected. This indicated that there had to be almost no further deaths in the intervention group for a significance level of 5% to be reached.

The CDP experience also warns against the dangers of stopping too soon [7, 51]. In the early months of the study, clofibrate appeared to be beneficial, with the significance level reaching or exceeding 5% on five monitoring occasions (Fig. 16.3). However, because of the repeated testing issue described earlier in this chapter, the decision was made to continue the study and closely monitor the results. The early difference was not maintained, and at the end of the trial the drug showed no benefit over placebo. It is notable that the mortality curves shown in Fig. 16.4 do not suggest

the wide swings observed in the interim analyses shown in Fig. 16.3. The fact that participants were entered over a period of time and thus had various lengths of follow-up at any given interim analysis, explains the difference between the two types of analyses. (See Chap. 15 for a discussion of survival analysis.). Pocock [52] also warns about the dangers of terminating trials too early for benefit, reflecting on a systematic review of trials stopped early [55]. At an early interim analysis, the Candesartan in Heart failure Assessment of Reduction in Mortality and Morbidity (CHARM) trial [65] had a 25% mortality benefit ($p<0.001$) from candesartan compared to a placebo control, but for a variety of reasons the trial continued and found after a median of 3 years of follow-up only a 9% nonsignificant difference in mortality. Continuing the trial revealed that the early mortality benefit was probably exaggerated and allowed other long-term intervention effects to be discovered. In general, trials stopped early for benefit often do not report in sufficient detail the rationale for early termination and often show implausibly large intervention effects based on only a small number of events [56]. This phenomenon is well recognized [57]. Thus, while there are sound ethical reasons to terminate trials early because of benefit, these decisions must be cautioned by our experience with early trends not being reliable or sustainable. Nevertheless, there is a natural tension between getting the estimate of treatment benefit precise and allowing too many participants to be exposed to the inferior intervention [56]. Statistical methods to be described later are useful as guides but not adequate as rules and the best approach based on experience is to utilize a properly constituted monitoring committee, charged with weighing the benefits and risks of early termination.

The Nocturnal Oxygen Therapy Trial was a randomized, multicenter clinical trial comparing two levels of oxygen therapy in people with advanced chronic obstructive pulmonary disease [66, 67]. While mortality was not considered as the primary outcome in the design, a strong mortality difference emerged during the trial, notably in one particular subgroup. Before any decision was made, the participating clinical centers were surveyed to ensure that the mortality data were as current as possible. A delay in reporting mortality was discovered and when all the deaths were considered, the trend disappeared. The earlier results were an artifact caused by incomplete mortality data. Although a significant mortality difference ultimately emerged, the results were similar across subgroups.

Early termination of a subgroup can be especially error prone if not done carefully. Peto and colleagues [68] have illustrated the danger of subgroup analysis by reporting that treatment benefit in ISIS-2 did not apply to individuals born during a certain astrologic sign. Nevertheless, treatment benefits may be observed in subgroups which may be compelling. An AIDS trial conducted by the AIDS Clinical Trial Research Group (ACTG), ACTG-019 [4, 5, 33] indicated that zidovudine (AZT) led to improved outcome in participants who had a low laboratory value (CD4 cell counts under 500 – which is a measure of poor immune response). The results were not significant for participants with a higher CD4 value. Given previous experience with this drug, and given the unfavorable prognosis for untreated AIDS patients, the trial was stopped early for benefit in those with the low CD4 cell count but continued in the rest of the participants.

A scientific and ethical issue was raised in the Diabetic Retinopathy Study, a randomized trial of 1,758 participants with proliferative retinopathy [69, 70]. Each participant had one eye randomized to photocoagulation and the other to standard care. After 2 years of a planned 5 year follow-up, a highly significant difference in the incidence of blindness was observed (16.3% vs. 6.4%) in favor of photocoagulation [71]. Since the long-term efficacy of this new therapy was not known, the early benefit could possibly have been negated by subsequent adverse reactions. After much debate, the monitoring committee decided to continue the trial, publish the early results, and allow any untreated eye at high risk of blindness to receive photocoagulation therapy [72]. In the end, the early treatment benefit was sustained over a longer follow-up, despite the fact that some of the eyes randomized to control received photocoagulation. Furthermore, no significant long-term adverse effect was observed.

The Beta-Blocker Heart Attack Trial provided another example of early termination [73, 74]. This randomized placebo control trial enrolled over 3,800 participants with a recent myocardial infarction to evaluate the effectiveness of propranolol in reducing mortality. After an average of a little over 2 years of a planned 3 year follow-up, a mortality difference was observed, as shown in Fig. 16.5. The results

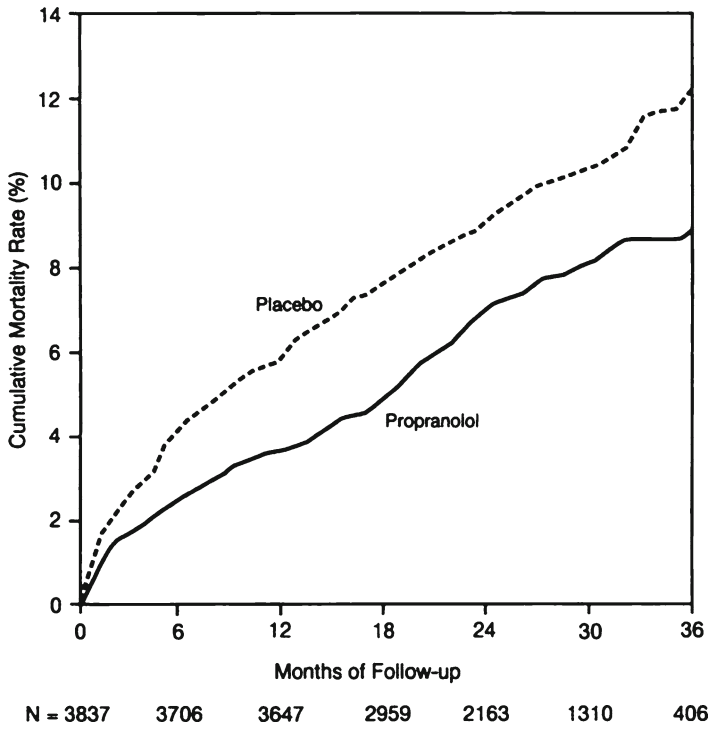


Fig. 16.5 Cumulative mortality curves comparing propranolol and placebo in the Beta-Blocker Heart Attack Trial [74]

were statistically significant, allowing for repeated testing, and would, with high probability, not be reversed during the next year [74]. The data monitoring committee debated whether the additional year of follow-up would add valuable information. It was argued that there would be too few events in the last year of the trial to provide a good estimate of the effect of propranolol treatment in the third and fourth year of therapy. Thus, the committee decided that prompt publication of the observed benefit was more important than waiting for the marginal information yet to be obtained. This trial was one of the early trials to implement group sequential monitoring boundaries discussed later in this chapter and will be used as an example to illustrate the method.

Another example of using sequential monitoring boundaries is found in chronic heart failure trials that evaluated different beta blockers. Common belief had been that administering a beta-blocker drug to a heart failure patient would cause harm, not benefit. Fortunately, early research suggested this belief may have been in error and ultimately four well designed trials were conducted to evaluate the risks and benefits. Three trials were terminated early because of beneficial intervention effect on mortality of 30–35% [75–77]. The fourth trial [78] did not go to completion in part due to the fact that the other three trials had already reported substantial benefits. Details of monitoring in one of the trials, the Metoprolol CR/XL Randomized Trial In Chronic Heart Failure (MERIT-HF) are discussed more fully later.

Some trials of widely used intervention have also been stopped early due to adverse events. One classic example comes from the treatment of arrhythmias following a heart attack. Epidemiological data showed an association between the presence of irregular ventricular heartbeats or arrhythmias and the incidence of sudden death, presumably due to serious arrhythmias. Drugs were developed that suppressed such arrhythmias and they became widely used after approval by the drug regulatory agency for that indication. The Cardiac Arrhythmia Suppression Trial (CAST) was a multicenter randomized double blind placebo-controlled trial evaluating the effects of three such drugs (encainide, flecainide, moricizine) on total mortality and sudden death [79]. Statistical procedures used in CAST to address the repeated testing problem [80, 81] are described later in the chapter. However, the encainide and flecainide arms of the trial were terminated after only 15% of the expected mortality events observed because of an adverse experience (63 deaths in the two active arms vs. 26 deaths in the corresponding placebo arms).

At the first monitoring committee review, the mortality trend in CAST began to appear but the number of events was relatively small [81]. Because the monitoring committee decided no definitive conclusion could be reached on the basis of so few events, it elected to remain blinded to the treatment assignment. However, before the next scheduled meeting, the statistical center alerted the committee that the trends continued and were now nearing the CAST monitoring criteria for stopping. In a conference call meeting, the monitoring committee became unblinded and learned that the trends were in the unexpected direction, that is, toward harm from the active treatment. A number of confirmatory and exploratory analyses were requested by the monitoring committee and a meeting was held a few weeks later to discuss fully these unexpected results. After a thorough review, the monitoring

committee recommended immediate termination of the encainide and flecainide portions of the trial [81]. Results were consistent across outcome variables and participant subgroups, and no biases could be identified which would explain these results. The third arm (moricizine) continued since there were no convincing trends at that time, but it too was eventually stopped due to adverse experiences [82]. The CAST experience points out that monitoring committees must be prepared for the unexpected and that large trends may emerge quickly. Even in this dramatic result, the decision was not simple or straightforward. Many of the issues discussed earlier were covered thoroughly before a decision was reached [81].

Not all negative trends emerge as dramatically as in the CAST. Two other examples are provided by trials in congestive heart failure. Yearly mortality from severe congestive heart failure is approximately 40%. The Prospective RandOmized MIlrinone Survival Evaluation (PROMISE) [35] and the Prospective RandOmized FlosequInan Longevity Evaluation (PROFILE) [36] trials evaluated inotropic agents (milrinone and flosequinone). Both of these drugs had been approved by regulatory agencies for use on the basis of improved exercise tolerance, which might be considered a surrogate response for survival. PROMISE and PROFILE were randomized placebo controlled trials comparing mortality outcomes. Both trials were unexpectedly terminated early due to statistically significant harmful mortality results, even after adjusting for repeated testing of these data. Because severe heart failure has a high mortality rate and the drugs were already in use, it was a difficult decision how long and how much evidence was needed to decide that the intervention was not helpful but was in fact harmful. In both trials, the monitoring committees allowed results to achieve statistical significance since a negative, but nonsignificant trend might have been viewed as evidence consistent with no effect on mortality.

The PROMISE and PROFILE experiences illustrate the most difficult of the monitoring scenarios, the emerging negative trend, but they are not unique [83–87]. Trials with persistent nonsignificant negative trends may have no real chance of reversing and indicating a benefit from intervention. In some circumstances, that observation may be sufficient to end the trial since if a result falls short of establishing benefit, the intervention would not be used. For example a new expensive or invasive intervention would likely need to be more effective than a standard intervention to be used. In other circumstances, a neutral result may be important, so a small negative trend, still consistent with a neutral result, would argue for continuation. If a treatment is already in clinical use on the basis of other indications, as in the case of the drugs used in PROMISE and PROFILE, an emerging negative trend may not be sufficient evidence to alter clinical practice. If a trial terminates early without resolving convincingly the harmful effects of an intervention, that intervention may still continue to be used. This practice would put future patients at risk, and perhaps even participants in the trial as they return to their usual healthcare system. In that case, the investment of participants, investigators, and sponsors would not have resolved an important question. There is a serious and delicate balance between the responsibility to safeguard the participants in the trial and the responsibility for all concurrent and future patients [83].

Trials may continue to their scheduled termination even though interim results are very positive and persuasive [88] or the intervention and control data are so similar that almost surely no significant results will emerge [89–92]. In one study of antihypertensive therapy, early significant results did not override the need for getting long-term experience with an intensive intervention strategy [88]. Another trial [90] implemented approaches to reduce cigarette smoking, change diet to lower cholesterol, and used antihypertensive medications to lower blood pressure in order to reduce the risk of heart disease. Although early results showed no trends, it was also not clear how long intervention needed to be continued before the applied risk factor modifications would take full effect. It was argued that late favorable results could still emerge. In fact, they did, though not until some years after the trial had ended [92]. In a trial that compared medical and surgical treatment of coronary artery artherosclerosis, the medical care group had such a favorable survival experience that there was little room for improvement by immediate coronary artery bypass graft intervention [91].

The Women's Health Initiative (WHI) was one of the largest and most complex trials ever conducted, certainly in women [93, 94]. This partial factorial trial evaluated three interventions in postmenopausal women: (1) hormone replacement therapy (HRT), (2) a low fat diet, and (3) calcium and vitamin D supplementation. Each intervention, in principle, could affect multiple organ systems, each with multiple outcomes. For example, HRT was being evaluated for its affect on cardiovascular events such as mortality and fatal and non-fatal myocardial infarction. HRT can also affect bone density, the risk of fracture, and breast cancer. The HRT component was also stratified into those with an intact uterus, who received both estrogen and progestin, and those without a uterus who received estrogen alone. The estrogen–progestin arm was terminated early due to increases in deep vein thrombosis, pulmonary embolism, stroke, and breast cancer and a trend toward increased heart disease as shown in Fig. 16.6a although there was a benefit in bone fracture as expected [93]. There was no observed difference in total mortality or the overall global index, the composite outcome defined in the protocol, as shown in Fig. 16.6b. The WHI is an excellent example of the challenges of monitoring trials with composite outcomes where component trends are not consistent. In such cases, the most important or most clinically relevant component may have to dominate in the decision process, even if not completely specified in the protocol or the monitoring committee charter. Later, the WHI estrogen-alone arm was also terminated, primarily due to increased pulmonary embolus and stroke, though there was no difference in myocardial infarction or total mortality [94]. The formal monitoring process had to account for multiple interventions, multiple outcomes and repeated testing.

A heart failure trial evaluating the drug tezosentan used a stopping criterion that included futility [95]. That is, when there was less than a 10% chance of having a positive beneficial result, the monitoring committee was to alert the investigators and sponsors and recommend termination. In fact, at about two-thirds of the way into the trial, a slightly negative trend was sufficient to make any chance of a beneficial result unlikely and the trial was terminated.

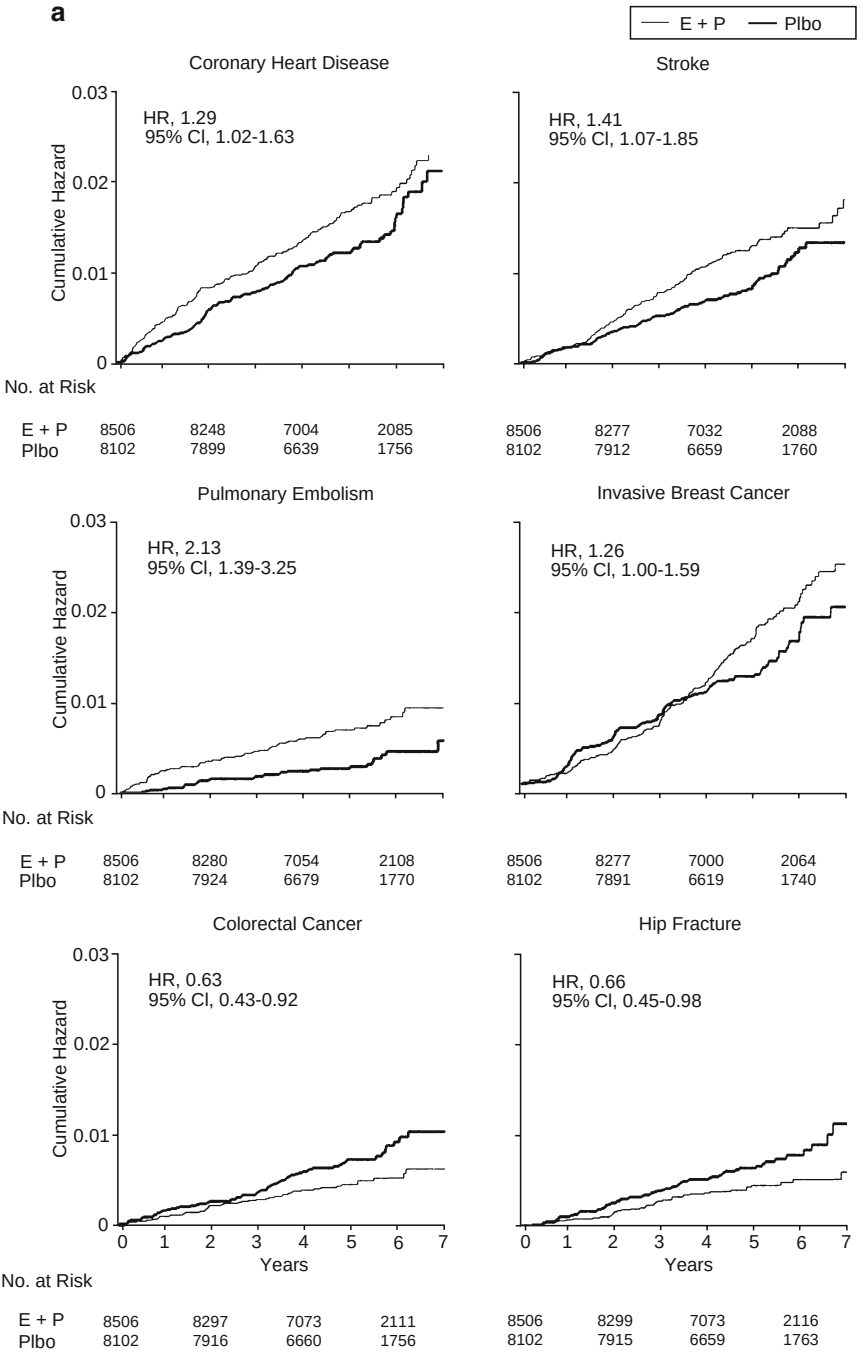


Fig. 16.6 (**a**) WHI Kaplan–Meier estimates of cumulative hazards for selected clinical outcomes HR = hazard ratio [93]

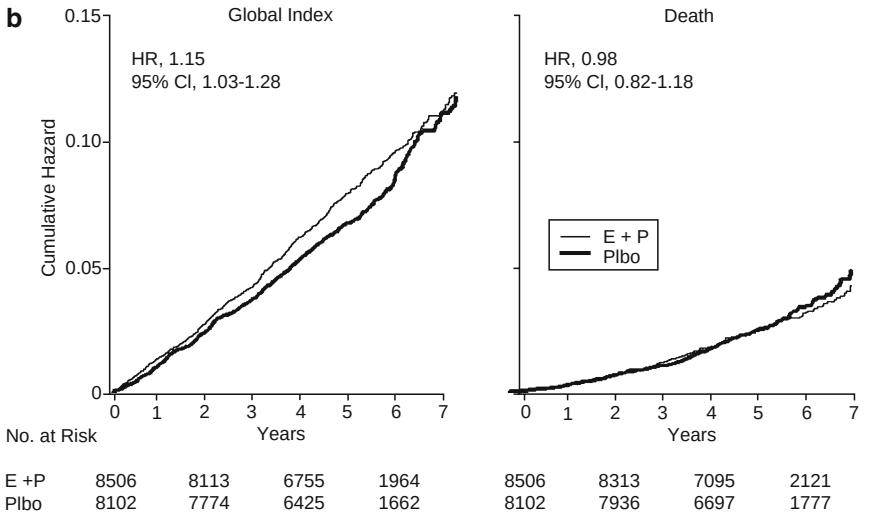


Fig 16.6 (continued) (**b**) WHI Kaplan–Meier estimates of cumulative hazards for global index and death HR = hazard ratio [93]

In some instances, a trial may be terminated because the hypothesis being tested has been convincingly answered by other ongoing trials. This was the case with trials evaluating warfarin in the treatment of atrial fibrillation [96]. Between 1985 and 1987, five trials were launched to evaluate warfarin to prevent strokes in participants with atrial fibrillation. Three of the trials were terminated early by 1990, reporting significant reductions in embolic complications. One of the remaining trials was also terminated early, largely due to the ethical aspects of continuing trials when the clinical question being tested has already been answered. The window of opportunity to further evaluate the intervention had closed.

In all of these studies, the decisions were difficult and involved many analyses, thorough review of the literature, and an understanding of the biological processes. As described above, a number of questions must be answered before serious consideration should be given to early termination. As noted elsewhere, the relationship between clinical trials and practice is very complex and this complexity is evident in the monitoring process [97, 98].

Decision to Extend a Trial

The issue about extending a trial beyond the original sample size or planned period of follow-up may arise. Suppose the mortality rate over a 2-year period in the control group is assumed to be 40%. (This estimate may be based on data from another trial involving a similar population.) Also specified is that the sample size should be large enough to detect a 25% reduction due to the intervention, with a

two-sided significance level of 5% and a power of 90%. The total sample size is, therefore, approximately 960. However, say that early in the study, the mortality rate in the control group appears somewhat lower than anticipated; or closer to 30%. This difference may result from a change in the study population, selection factors in the trial, or new concomitant therapies. If no design changes are made, the intervention would have to be more effective (30% reduction rather than 25%) for the difference between groups to be detected with the same power. Alternatively, the investigators would have to be satisfied with approximately 75% power of detecting the originally anticipated 25% reduction in mortality. If it is unreasonable to expect a 30% benefit and if a 75% power is unacceptable, the design needs modification. Given the lower control group mortality rate, approximately 1,450 participants would be required to detect a 25% reduction in mortality, while maintaining a power of 90%. Another option is to extend the length of follow-up, which would increase the overall event rate. A combination of these two approaches can also be tried.

The concept of adaptive designs has already been discussed in Chap. 5. Adaptive designs can be used in trials with overall lower event rates or increased variability, or when emerging trends are smaller than planned for but yet of clinical interest. Modifying the design once the trial is underway due to lower event rates or increased variability is rather straightforward. In a trial of antenatal steroid administration [99], the incidence of infant respiratory distress in the control group was much less than anticipated. Early in the study, the investigators decided to increase the sample size by extending the recruitment phase. In another trial, the protocol specifically called for increasing the sample size if the control group event rate was less than assumed [100]. As described in the sample size chapter, power is the probability of detecting a treatment effect if there truly is an effect. This probability is computed at the beginning of the trial during the design phase. The design goal is to set this probability at a range from 0.80 to 0.95 with an appropriate sample size. Sometimes this probability, or power, is referred to as "unconditional power" to distinguish it from "conditional power" to be described in more detail later in this chapter. Adjustments to sample size based on overall event rates or variability estimates can preserve the power (or unconditional power). No account of emerging trends is used in this recalculation.

The issue of whether the control group event rate or the overall event rate should be used in this sample size reassessment must be considered. It might seem intuitive that the emerging control group event rate should be used since it was the estimated control group rate that was initially used in the sample size calculation, as described in Chap. 8. However, to reveal the control group rate to the investigators may unblind the emerging trend if they are also aware of the overall number of events. The use of the overall event rate would avoid this potential problem. Additionally, there are statistical arguments that under the null hypothesis, the overall rate is the more appropriate one to use because it is likely to be more stable, particularly if the sample size re-estimation is done early in the trial. We prefer to use the overall event rate, but in either case, this must be decided while the protocol and data monitoring procedures are being developed.

However, modifying the design based on emerging trends is more complicated (see Chap. 5) and will be discussed in more technical detail later in this chapter. Despite the statistical literature for different approaches [101–104] and some criticism [105, 106], only a few applications of this type of adaptive design have been utilized. One such trial is the African-American Heart Failure Trial (A-HeFT) [107], a trial in African Americans with advanced heart failure using a combination of two established drugs. The primary outcome consisted of a weighted score of death, hospitalization, and quality of life. Mortality was among the secondary outcomes. The trial utilized an adaptive design [101] that required the monitoring committee to assess variability of this novel primary outcome and the emerging trend to make sample size adjustment recommendations to the trial leaders. The reason for the adaptive design was that little previous data were available for this combined outcome so estimates of variability were not adequate to compute a reliable sample size. Little experience with the outcome also limited the assessment of potential drug effect on this outcome. A group sequential boundary was established using a Lan–DeMets alpha spending function of the O'Brien–Fleming type, described later in this chapter, for monitoring benefit or harm for the composite outcome. This adaptive procedure was followed as planned and the sample size was increased from 800 to 1,100. Meanwhile, the monitoring committee was observing a mortality trend favoring the combination drug but there was no sequential monitoring plan prespecified for this outcome. The monitoring committee elected to utilize the same sequential boundary specified for the primary composite outcome to monitor mortality. Although not ideal while the trial was ongoing, it was done before the mortality difference became nominally significant. At the last scheduled meeting of the monitoring committee, the difference was nominally significant at the 0.05 level but had not crossed the sequential boundary. The committee decided to conduct an additional review of the data. At that additional review, the mortality difference was nominally significant ($p=0.01$) and had crossed the sequential O'Brien–Fleming boundary. The committee recommended early termination both because of a significant mortality benefit and a primary outcome that was nominally significant, along with a consistency across the components of the composite outcome and relevant subgroups.

While the statistical methods for adaptive designs based on emerging trends to reset the sample size exist, the use of these methods is still evolving. A more technical discussion of specific trend adaptive designs is provided later in this chapter. One concern is whether the application of the prespecified algorithm, according to the statistical plan, may reveal information about the size and direction of the emerging trend to those blind to the data. We are aware that these algorithms can be "reverse engineered" and a reasonable estimate of the emerging trend can be obtained. We know of no example to date where this revelation has caused a problem but in principle this could cause bias in participant selection or recruitment efforts or even participant assessment. Thus, mechanisms for implementation of trend adaptive trials are needed that protect the integrity of the trial.

Another approach that has been used [35, 36] is to fix the target of the trial to be a specified number of events in the control group or for the total number. If event

rates are low, it may take longer follow-up per participant or more randomized participants, or both, to reach the required number of events. In any case, the target is the number of events. In the above situations, only data from the control group or the combined groups are used. No knowledge of what is happening in the intervention group is needed. However, if the intervention group results are not used in the recalculations, then an increase in sample size could be recommended when the observed difference between the intervention and control groups is actually larger than originally expected. Thus, in the hypothetical example described above, if early data really did show a 30% benefit from intervention, an increased sample size might not be needed to maintain the desired power of 90%. For this reason, one would not like to make a recommendation about extension without also considering the observed effect of intervention. Computing conditional power is one way of incorporating these results. Conditional power is the probability that the test statistic will be larger than the critical value, given that a portion of the statistic is already known from the observed data and described later in this chapter. As in other power calculations, the assumed true difference in response variables between groups must be specified. When the early intervention experience is better than expected, the conditional power will be large. When the intervention is doing worse than anticipated, the conditional power will be small. The conditional power concept utilizes knowledge of outcome in both the intervention and control groups and is, therefore, controversial. Nevertheless, the concept attempts to quantify the decision to extend.

Towards the scheduled end of a trial, the investigator may find that he has nearly statistically significant results. He may be tempted to extend or expand the trial in an effort to make the test statistic significant. Such a practice is not recommended. A strategy of extending assumes that the observed relative differences in rates of response will continue. The observed differences that are projected for a larger sample may not hold. In addition, because of the multiple testing issue and the design changes, the significance level should be adjusted to a smaller value. However, appropriate adjustments in the significance level to account for the design changes may not easily be determined. Since a more extreme significance level should be employed, and since future responses are uncertain, extension may leave the investigator without the expected benefits.

Whatever adjustments are made to either sample size or the length of follow-up, they should be made as early in the trial as possible or as part of a planned adaptive design strategy. Early adjustments would diminish the criticism that the investigator or the monitoring committee waited until the last minute to see whether the results would achieve some prespecified significance level before changing the study design.

Statistical Methods Used in Monitoring

In this section, some statistical methods currently available for monitoring the accumulating data in a clinical trial will be reviewed. The methods address whether the trial should be terminated early or continued to its planned termination. No single

statistical test or monitoring procedure ought to be used as a strict rule for decision-making, but rather as one piece of evidence to be integrated with other evidence [3–8]. Most methods are very specific in their applications. Therefore, it is difficult to make a single recommendation about which should be used. However, the following methods, when applied appropriately, can be useful guides in the decision-making process.

Classical sequential methods, a modification generally referred to as group sequential methods, and curtailed testing procedures are discussed below for data monitoring. Other approaches are also briefly considered. Classical sequential methods are given more mathematical attention in several articles and texts [108–115].

Classical Sequential Methods

The aim of the classical sequential design is to minimize the number of participants that must be entered into a study. The decision to continue to enroll participants depends on results from those already entered. Most of these sequential methods assume that the response variable outcome is known in a short time relative to the duration of the trial. Therefore, for many trials involving acute illness, these methods are applicable. For studies involving chronic diseases, classical sequential methods have not been as useful. Although the sequential approaches have design implications, we have delayed discussing any details until this chapter because they really focus on monitoring accumulating data. Even if, during the design of the trial, consideration were not given to sequential methods, they could still be used to assist in the data monitoring or the decision-making process. Detailed discussions of classical sequential methods are given, for example, by Armitage [110], Whitehead [113], and Wald [108].

The sequential analysis method as originally developed by Wald [108] and applied to the clinical trial by others such as Armitage [48, 110] involves repeated testing of data in a single experiment. The method assumes that the only decision to be made is whether the trial should continue or be terminated because one of the groups is responding significantly better, or worse, than the other. This classical sequential decision rule is called an "open plan" by Armitage [110] because there is no guarantee of when a decision to terminate will be reached. Strict adherence to the "open plan" would mean that the study could not have a fixed sample size. Very few clinical trials use the "open" or classical sequential design. The method also requires data to be paired, one observation from each group. In many instances, the pairing of participants is not appealing because the paired participants may be very different and may not be "well matched" in important prognostic variables. If stratification is attempted in order to obtain better matched pairs, each stratum with an odd number of participants would have one unpaired participant. Furthermore, the requirement to monitor the data after every pair may be impossible or unnecessary for many clinical trials. Silverman and colleagues [116] used an "open plan" in a trial of the effects of humidity on survival in infants with low birth weight. At the end of 36 months, 181 pairs of infants had been enrolled; 52 of the pairs had a

discrepant outcome. Nine infants were excluded because they were un-matched and 16 pairs were excluded because of a mismatch. The study had to be terminated without a clear decision because it was no longer feasible to continue the trial. This study illustrates the difficulties inherent in the classical sequential design.

Armitage [48] introduced the restricted or "closed" sequential design to assure that a maximum limit is imposed on the number of participants ($2N$) to be enrolled. As with the "open plan," the data must be paired using one observation from each study group. Criteria for early termination and rejection of no treatment effect are determined so that the design has specified levels of significance and power (α and $1-\beta$). The restricted plan was used in a comparison of two interventions in patients with ulcerative colitis [117]. In that trial, the criterion for no treatment effect was exceeded, demonstrating short-term clinical benefit of corticosteroids over sulphasalazine therapy. This closed design was also used in an acute leukemia trial, comparing 6-mercaptopurine with placebo (CALGB) [118]. This trial was terminated early, with the statistic comparing remission rates crossing the sequential boundary for benefit after 21 pairs of patients.

Another solution to the repeated testing problem, called "repeated significance tests," was proposed by McPherson and Armitage [119] and also described by Armitage [110]. Although different theoretical assumptions are used, this approach has features similar to the restricted sequential model. That is, the observed data must be paired, and the maximum number of pairs to be considered can be fixed. Other modifications to the Armitage restricted plan [120–122] have also been proposed.

The methods described above can in some circumstances be applied to interim analyses of censored survival data [122–131]. If participants simultaneously enter a clinical trial and there is no loss to follow-up, information on interim analyses is said to be "progressively censored." Sequential methods for this situation have been developed using, for example, modified rank statistics. In fact, most participants are not entered into a trial simultaneously, but in a staggered fashion. That is, participants enter over a period of time with events of interest occurring after that, subject to an independent censoring process. The log-rank statistic, described in Chap. 15, may also be used in this situation.

The classical sequential approach has not been widely used, even in clinical trials where the time to the event is known almost immediately. One major reason perhaps is the requirement of analysis after every pair of outcomes or events. For many clinical trials, this is not necessary or even feasible if the data are monitored by a committee which has regularly scheduled meetings. In addition, classical sequential boundaries require an alternative hypothesis to be specified, a feature not demanded by conventional statistical tests for the rejection of the null hypothesis.

Group Sequential Methods

Because of limitations with classical sequential methods, other approaches to the repeated testing problem have been proposed. Ad hoc rules have been suggested

that attempt to ensure a conservative interpretation of interim results. One such method is to use a critical value of 2.6 at each interim look as well as in the final analyses [7]. Another approach [15, 132] referred to as the Haybittle–Peto procedure, favors using a large critical value, such as $Z_i = +3.0$, for all interim tests ($i < K$). Then any adjustment needed for repeated testing at the final test ($i = K$) is negligible and the conventional critical value can be used. These methods are ad hoc in the sense that no precise Type I error level is guaranteed. They might, however, be viewed as precursors of the more formal procedures to be described below.

Pocock [133–135] modified the repeated testing methods of McPherson and Armitage [119] and developed a group sequential method for clinical trials which avoids many of the limitations of classical methods. He discusses two cases of special interest; one for comparing two proportions and another for comparing mean levels of response. Pocock's method divides the participants into a series of K equal-sized groups with $2n$ participants in each, n assigned to intervention and n to control. K is the number of times the data will be monitored during the course of the trial. The total expected sample size is $2nK$. The test statistic used to compare control and intervention is computed as soon as data for the first group of $2n$ participants are available, and recomputed when data from each successive group become known. Under the null hypothesis, the distribution of the test statistic, Z_i, is assumed to be approximately normal with zero mean and unit variance, where i indicates the number of groups ($i \leq K$) which have complete data. This statistic Z_i is compared to the stopping boundaries, $\pm ZN_K$ where ZN_K has been determined so that for up to K repeated tests, the overall (two sided) significance level for the trial will be α. For example, if $K = 5$ and $\alpha = 0.05$ (two-sided), $ZN_K = 2.413$. This critical value is larger than the critical value of 1.96 used in a single test of hypothesis with $\alpha = 0.05$. If the statistic Z_i falls outside the boundaries on the "i"-th repeated test, the trial should be terminated, rejecting the null hypothesis. If the statistic falls inside the boundaries, the trial should be continued until $i = K$ (the maximum number of tests). When $i = K$, the trial would stop and the investigator would "accept" H_0.

O'Brien and Fleming [136] also discuss a group sequential procedure. Using the above notation, their stopping rule compares the statistic Z_i with $Z^* \sqrt{(K / i)}$ where Z^* is determined so as to achieve the desired significance level. For example, if $K = 5$ and $\alpha = 0.05$, $Z^* = 2.04$. If $K \leq 5$, Z^* may be approximated by the usual critical values for the normal distribution. One attractive feature is that the critical value used at the last test ($i = K$) is approximately the same as that used if a single test were done.

In Fig. 16.7, boundaries for the three methods described are given for $K = 5$ and $\alpha = 0.05$ (two-sided). If for $i < 5$ the test statistic falls outside the boundaries, the trial is terminated and the null hypothesis rejected. Otherwise, the trial is continued until $i = 5$, at which time the null hypothesis is either rejected or "accepted". The three boundaries have different early stopping properties. The O'Brien–Fleming model is unlikely to lead to stopping in the early stages. Later on, however, this procedure leads to a greater chance of stopping prior to the end of the study than the other two. Both the Haybittle–Peto and the O'Brien–Fleming boundaries avoid

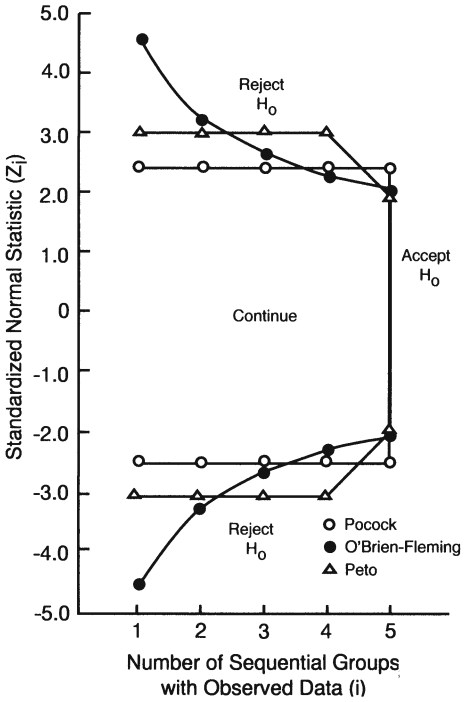


Fig. 16.7 Three group sequential stopping boundaries for the standardized normal statistic (Z_i) for up to five sequential groups with two-sided significance level of 0.05 [93]

the awkward situation of accepting the null hypothesis when the observed statistic at the end of the trial is much larger than the conventional critical value (i.e., 1.96 for a two-sided 5% significance level). If the observed statistic in Fig. 16.7 is 2.3 when $i=5$, the result would not be significant using the Pocock boundary. The large critical values used at the first few analyses for the O'Brien–Fleming boundary can be adjusted to some less extreme values (e.g., 3.5) without noticeably changing the critical values used later on, including the final value.

Many monitoring committees wish to be somewhat conservative in their interpretation of early results because of the uncertainties discussed earlier and because a few additional events can alter the results substantially. Yet, most investigators would like to use conventional critical values in the final analyses, not requiring any penalty for interim analyses. This means that the critical value used in a conventional fixed sample methods would be the same for that used in a sequential plan, resulting in no increase in sample size. With that in mind, the O'Brien–Fleming model has considerable appeal, perhaps with the adjusted or modified boundary as described. The group sequential methods have an advantage over the classical methods in that the data do not have to be continuously tested and individual participants do not have to be "paired." This concept suits the data review activity of most large clinical trials where monitoring committees meet periodically.

Furthermore, in many trials constant consideration of early stopping is unnecessary. Pocock [133–135] discusses the benefits of the group sequential approach in more detail and other authors describe variations [137–141].

In many trials, participants are entered over a period of time and followed for a relatively long period. Frequently, the primary outcome is time to some event. Instead of adding participants between interim analyses, new events are added. As discussed in Chap. 15, survival analysis methods could be used to compare the experience of the intervention and the control arms. Given their general appeal, it would be desirable to use the group sequential methods in combination with survival analyses. It has been established for large studies that the log-rank or Mantel–Haenszel statistic [142–147] can be used. Furthermore, even for small studies, the log-rank procedure is still quite robust. The Gehan, or modified Wilcoxon test [148, 149], as defined in Chap. 15 cannot be applied directly to the group sequential procedures. A generalization of the Wilcoxon procedure for survival data, though, is appropriate [150] and the survival methods of analyses can in general terms be applied in group sequential monitoring. Instead of looking at equal-sized participant groups, the group sequential methods described strictly require that interim analyses should be done after an additional equal number of events have been observed. Since monitoring committees usually meet at fixed calendar times, the condition of equal number of events might not be met exactly. However, the methods applied under these circumstances are approximately correct [151] if the increments are not too disparate. Other authors have also described the application of group sequential methods to survival data [152–155].

Interim log-rank tests in the Beta-Blocker Heart Attack Trial [74] were evaluated using the O'Brien–Fleming group sequential procedure [136]. Seven meetings had been scheduled to review interim data. The trial was designed for a two-sided 5% significance level. These specifications produce the group sequential boundary shown in Fig. 16.8. In addition, the interim results of the log-rank statistic are also shown for the first six meetings. From the second analysis on, the conventional significance value of 1.96 was exceeded. Nevertheless, the trial was continued. At the sixth meeting, when the O'Brien–Fleming boundary was crossed, a decision was made to terminate the trial with the final mortality curves as seen in Fig. 16.5. It should be emphasized that crossing the boundary was not the only factor in this decision.

Flexible Group Sequential Procedures: Alpha Spending Functions

While the group sequential methods described are an important advance in data monitoring, the Beta-blocker Heart Attack Trial (BHAT) [74] experience suggested two limitations. One was the need to specify the number K of planned

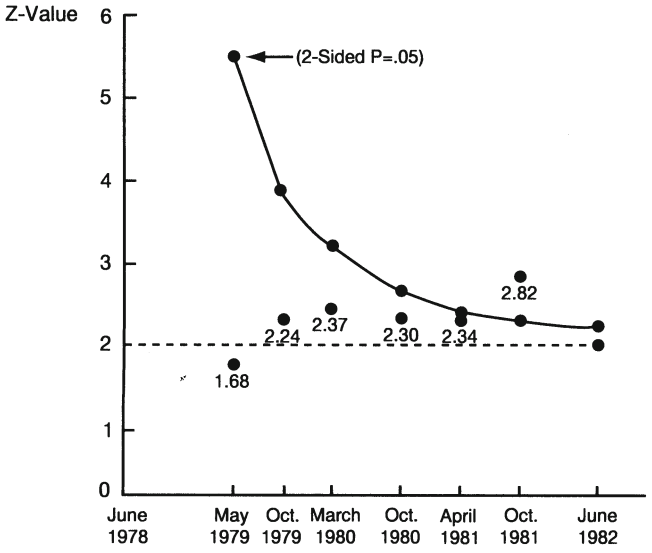


Fig. 16.8 Six interim log-rank statistics plotted for the time of data monitoring committee meetings with a two-sided O'Brien–Fleming significance level boundary in the Beta-Blocker Heart Attack Trial. *Dashed line* represents $Z = 1.96$ [74]

interim analyses in advance. The second was the requirement for equal numbers of either participants or events between each analysis. This also means that the exact time of the interim analysis must be pre-specified. As indicated in the BHAT example, the numbers of deaths between analyses were not equal and exactly seven analyses of the data had been specified. If the monitoring committee had requested an additional analysis between the fifth and sixth scheduled meetings, the O'Brien–Fleming group sequential procedure would not have directly accommodated such a modification. Yet such a request could easily have happened. In order to accommodate the unequal numbers of participants or events between analyses and the possibility of larger or fewer numbers of interim analyses than pre-specified, flexible procedures that eliminated those restrictions were developed [156–163]. The authors proposed a so-called alpha spending function which allows investigators to determine how they want to "spend" the Type I error or alpha during the course of the trial. This function guarantees that at the end of the trial, the overall Type I error will be the prespecified value of α. As will be described, this approach is a generalization of the previous group sequential methods so that the Pocock [133] and O'Brien–Fleming [136] monitoring procedures become special cases.

We must first distinguish between calendar time and information fraction [160, 163]. At any particular calendar time t in the study, a certain fraction t^* of the total information is observed. That may be approximated by the fraction of participants randomized at that point, n, divided by the total number expected, N, or in survival

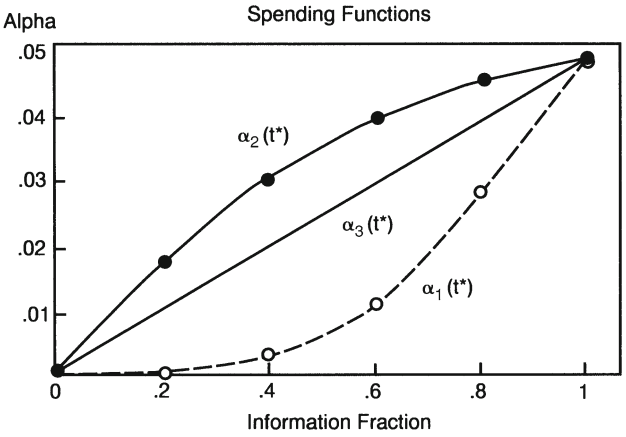


Fig. 16.9 Alpha-spending functions for $K=5$, two-sided $\alpha=0.05$ at information fractions $t^*=0.2$, 0.4, 0.6, 0.8, and 1.0 where $\alpha_1(t^*)\sim$O'Brien–Fleming; $\alpha_2(t^*)\sim$Pocock; $\alpha_3(t^*)\sim$uniform alpha spending functions [211]

studies, by the number of events observed already, d, divided by the total number expected D. Thus, the value for t^* must be between 0 and 1. The information fraction is more generally defined in terms of ratio of the inverse of the variance of the test statistic at the particular interim analysis and the final analysis. The alpha-spending function, $\alpha(t^*)$, determines how the pre-specified α is allocated at each interim analyses as a function of the information fraction. At the beginning of a trial, $t^*=0$ and $\alpha(t^*)=0$, while at the end of the trial, $t^*=1$ and $\alpha(t^*)=\alpha$. Alpha-spending functions that correspond to the Pocock and O'Brien–Fleming boundaries shown in Fig. 16.7 are indicated in Fig. 16.9 for a two-sided 0.05 α level and five interim analyses. These spending functions correspond to interim analyses at information fractions at 0.2, 0.4, 0.6, 0.8, and 1.0. However, in practice the information fractions need not be equally spaced. We chose those information fractions to indicate the connection between the earlier discussion of group sequential boundaries and the α spending function. The Pocock-type spending function allocates the alpha more rapidly than the O'Brien–Fleming type spending function. For the O'Brien–Fleming-type spending function at $t^*=0.2$, the $\alpha(0.2)$ is less than 0.0001 which corresponds approximately to the very large critical value or boundary value 4.56 in Fig. 16.7. At $t^*=0.4$, the amount of α which can be spent is $\alpha(0.4)-\alpha(0.2)$ which is approximately 0.0006, corresponding to the boundary value 3.23 in Fig. 16.7. That is, the difference in $\alpha(t^*)$ at two consecutive information fractions, t^* and t^{**} where t^* is less than t^{**}, $\alpha(t^{**})-\alpha(t^*)$, determines the boundary or critical value at t^{**}. Obtaining these critical values consecutively requires numerically integrating a distribution function similar to that for Pocock and is described elsewhere in detail [156]. Because these spending functions are only approximately equivalent to the Pocock or O'Brien–Fleming boundaries, the actually boundary values will be similar but not exactly the same. However, the practical differences are important. Programs are available for these calculations [164].

Many different spending functions can be specified. The O'Brien–Fleming $\alpha_1(t^*)$ and Pocock $\alpha_2(t^*)$ type spending functions are specified as follows:

$$\alpha_1(t^*) = 2 - 2\Phi\left(Z_{\alpha/2}/\sqrt{t^*}\right) \qquad \sim \text{O'Brien-Fleming}$$
$$\alpha_2(t^*) = \alpha \ln(1+(e-1)t^*) \qquad \sim \text{Pocock}$$
$$\alpha_3(t^*) = \alpha\, t^{*\theta} \qquad \text{for } \theta > 0$$

The spending function $\alpha_3(t^*)$ spends alpha uniformly during the trial for $\theta = 1$, at a rate somewhat between $\alpha_1(t^*)$ and $\alpha_2(t^*)$. Other spending functions have also been defined [165, 166].

The advantage of the alpha-spending function is that neither the number nor the time of the interim analyses needs to be specified in advance. Once the particular spending function is selected, the information fractions $t_1^*, t_2^*, \ldots$ determine the critical or boundary values exactly. In addition, the frequency of the interim analyses can be changed during the trial and still preserve the prespecified α level. Even if the rationale for changing the frequency is dependent on the emerging trends, the impact on the overall Type I error rate is almost negligible [167, 168]. These advantages give the spending function approach to group sequential monitoring the flexibility in analysis times that is often required in actual clinical trial settings [169]. It must be emphasized that no change of the spending function itself is permitted during the trial. Other authors have discussed additional aspects of this approach [44, 170, 171].

Applications of Group Sequential Boundaries

As indicated in the BHAT example [73, 74], the standardized logrank test can be compared to the standardized boundaries provided by the O'Brien–Fleming, Pocock, or α spending function approach. However, these group sequential methods are quite widely applicable for statistical tests which can be standardized with a normal distribution and independent increments of information between interim analyses. Besides logrank and other survival tests, comparisons of means, comparison of proportions [133, 172] and comparison of linear regression slopes [173–178] can be monitored using this approach. For means and proportions, the information fraction can be approximated by the ratio of the number of participants observed to the total expected. For regression slopes, the information fraction is best determined from the ratio of the inverse of the variance of the regression slope differences computed for the current and expected final estimate [174, 178]. Considerable work has extended the group sequential methodology to more general linear and nonlinear random effects models for continuous data and to repeated measure methods for categorical data [179]. Thus, for most of the statistical tests that would be applied to common primary outcome measures, the flexible group sequential methods can be used directly.

If the trial continues to the scheduled termination point, a p value is often computed to indicate the extremeness of the result. If the standardized statistical test exceeds the critical value, the p value would be less than the corresponding significance level. If a trial is terminated early or continues to the end with the standardized test exceeding or crossing the boundary value, a p value can also be computed [180]. These p values cannot be the nominal p value corresponding to the standardized test statistic. They must be adjusted to account for the repeated statistical testing of the outcome measure and for the particular monitoring boundary employed. Calculation of the p value is relatively straight forward with existing software packages [164].

Statistical tests of hypotheses are but one of the methods used to evaluate the results of a clinical trial. Once trials are terminated, either on schedule or earlier, confidence intervals (CIs) are often used to give some sense of the uncertainty in the estimated treatment or intervention effect. For a fixed sample study, CIs are typically constructed as

$$(\text{effect estimate}) \pm Z(\alpha)\,(\text{standard error of the estimate})$$

In the group sequential monitoring setting, this CI will be referred to as the naive estimate since it does not take into account the sequential aspects. In general, construction of CIs following the termination of a clinical trial is not as straightforward [181–194], but software exists to aid in the computations [164]. The major problem with naive CIs is that they may not give proper coverage of the unknown but estimated treatment effect. That is, the CIs constructed in this way may not include the true effect the proper number of times (e.g., 95%). For example, the width of the CI may be too narrow. Several methods have been proposed for constructing a more proper CI [181–194] typically ordering the possible outcomes in different ways. That is, a method is needed to determine if a treatment effect at one time is either more or less extreme than a difference at another time. None of the methods proposed appear to be universally superior but the ordering originally suggested by Siegmund [188] and adopted by Tsiatis et al. [181] appears to be quite adequate. In this ordering, any treatment comparison statistic which exceeds the group sequential boundary at one time is considered to be more extreme than any result which exceeds the sequential boundary at a later time. While construction of CIs using this ordering of possible outcomes can break down, the cases or circumstances are almost always quite unusual and not likely to occur in practice [189]. It is also interesting that for conservative monitoring boundaries such as the O'Brien–Fleming method, the naive CI does not perform that poorly, due primarily to the extreme early conservatism of the boundary [187]. While more exact CIs can be computed for this case, the naive estimate may still prove useful as a quick estimate to be recalculated later using the method described [181]. Pocock and Hughes [186] have suggested that the point estimate of the effect of the intervention should also be adjusted, since trials that are terminated early tend to exaggerate the size of the true treatment difference. Others have also pointed out the bias in the point estimate [182, 184]. Kim [184] suggested that an estimate of the median is less biased.

CIs can also be used in another manner in the sequential monitoring of interim data. At each interim analysis, a CI could be constructed for the parameter summarizing the intervention effect, such as differences in means, proportions, or hazard ratios. This is referred to as repeated confidence intervals (RCIs) [192–194]. If the RCI excludes a null difference, or no intervention effect, then the trial might be stopped claiming a significant effect. It is also possible to continue the trial unless the CI excluded not only no difference but also minimal or clinically unimportant differences. On the other hand, if all values of clinically meaningful treatment differences are ruled out or fall outside the CI, then that trial might be stopped claiming that no useful clinical effect is likely. Here, as for CIs following termination, the naive CI is not appropriate. Jennison and Turnbull [192, 193] have suggested one method for RCIs that basically inverts the group sequential test. That is, the CI has the same form as the naive estimate, but the coefficient is the standardized boundary value as determined by the spending function, for example. The RCI has the following form:

$$(\text{treatment difference}) \pm Z(k)\,(\text{standard error of difference})$$

where $Z(k)$ is the sequential boundary value at the kth interim analysis. For example, using the O'Brien–Fleming boundaries shown in Fig. 16.7, we would have a coefficient of 4.56 at $k=1$, $t_1^* = 0.2$ and 3.23 at $k=2$, $t_2^* = 0.4$. Used in this manner, the RCI and the sequential test of the null hypothesis will yield the same conclusions.

One particular application of the RCI is for trials whose goal is to demonstrate that two interventions or treatments are essentially "equivalent," that is, have an effect that is considered to be within a specified acceptable range and might be used interchangeably. As indicated in Chap. 5, clinicians might select the cheaper, less toxic or less invasive intervention if the effects were close enough. One suggestion for "close enough" or "equivalence" would be treatments whose effects are within 20% [195, 196]. Thus, RCIs that are contained within a 20% range would suggest that the results are consistent with this working definition of equivalence. For example, if the relative risks were estimated along with a RCI, the working range of equivalence would be from 0.8 to 1.2, where large values indicate inferiority of the intervention being tested. The trial would continue as long as the upper limit of the RCI exceeded 1.2 since we would not have ruled out a treatment worsening by 20% or more. Depending on the trial and the interventions, the trial might also continue until the lower limit of the RCI was larger than 0.8, indicating no improvement by 20% or greater.

As described in Chap. 5, there is a fundamental difference between an "equivalence" design and a noninferiority design. The former is a two-sided test, with the aim of establishing a narrow range of possible differences between the new intervention and the standard, or that any difference is within a narrow range. The noninferiority design aims to establish that the new intervention is no worse than the standard by some prespecified margin. It may be that the margins in the two designs are set to the same value. From a data monitoring point of view, both of these designs are best handled by sequential CIs [193]. As data emerge, the RCI takes

into consideration the event rate or variability, the repeated testing aspects, and the level of the CI. The upper and lower boundaries can address either the "equivalence" point of view or the noninferiority margin of indifference.

Asymmetric Boundaries

In most trials, the main purpose is to test whether the intervention is superior to the control. It is not always ethical to continue a study in order to prove, at the usual levels of significance, that the intervention is harmful relative to a placebo or standard control. This point has been mentioned by authors [197, 198] who discuss methods for group sequential designs in which the hypothesis to be tested is one-sided; that is, to test whether the intervention is superior to the control. They proposed retaining the group sequential upper boundaries of methods such as Pocock, Haybittle–Peto, or O'Brien–Fleming for rejection of H_0 while suggesting various forms of a lower boundary which would imply "acceptance" of H_0. One simple approach is to set the lower boundary at an arbitrary value of Z_i, such as −1.5 or −2.0. If the test statistic goes below that value, the data may be sufficiently suggestive of a harmful effect to justify terminating the trial. This asymmetric boundary attempts to reflect the behavior or attitude of members of many monitoring committees, who recommend stopping a study once the intervention shows a strong, but non-significant, trend in an adverse direction for major events. Emerson and Fleming [199] recommend a lower boundary for acceptance of the null hypothesis which allows the upper boundary to be changed in order to preserve the Type I error exactly. Work by Gould and Pecore [200] suggests ways for early acceptance of the null hypothesis while incorporating costs as well. For new interventions, trials might well be terminated when the chances of a positive or beneficial result seem remote (discussed in the next section). However, if the intervention arm is being compared to a standard but the intervention is already in widespread use, it may be important to distinguish between lack of benefit and harm [83]. For example, if the intervention is not useful for the primary outcome, and also not harmful, it may still have benefits such as on other clinical outcomes, quality of life, or fewer side effects, that would still make it a therapeutic option. In such cases, a symmetric boundary for the primary outcome might be appropriate.

An example of asymmetric group sequential boundaries is provided by the Cardiac Arrhythmia Suppression Trial (CAST). Two arms of the trial (encainide and flecainide, each vs. placebo) were terminated early using a symmetric two-sided boundary, although the lower boundary for harm was described as advisory by the authors [80, 81]. The third comparison (moricizine vs. placebo) continued. However, due to the experience with the encainide and flecainide arms, the lower boundary for harm was revised to be less stringent than originally, i.e., an asymmetric boundary was used [82].

MERIT-HF used a modified version of the Haybittle–Peto boundary for benefit, requiring a critical value near +3.0 and a similar but asymmetric boundary, close to a critical Z value of −2.5 for harm as shown in Fig. 16.10. In addition, at least 50% of

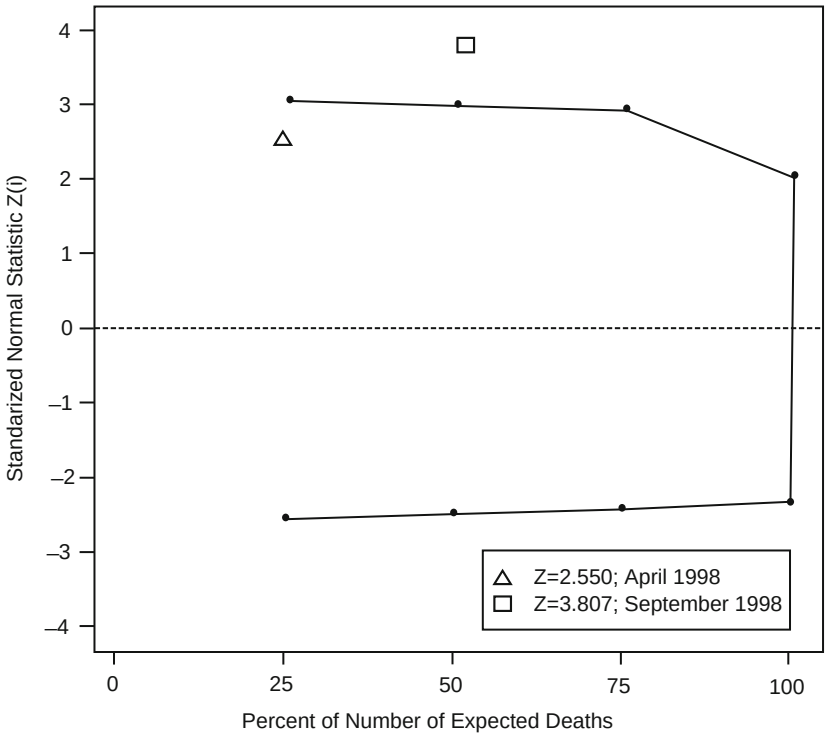


Fig. 16.10 MERIT-HF group sequential monitoring bounds for mortality

the designed person years of exposure should be observed before early termination could be recommended. The planned interim analyses to consider benefit were at 25, 50, and 75% of the expected target number of events. Because there was a concern that treating heart failure with a beta blocker might be harmful, the monitoring committee was required to evaluate safety on a monthly basis using the lower sequential boundary as a guide. At the 25% interim analyses, the statistic for the logrank test was +2.8, just short of the boundary for benefit. At the 50% interim analyses, the observed logrank statistic was +3.8, clearly exceeding the sequential boundary for benefit. It also met the desired person years of exposure as plotted in Fig. 16.10. Details of this experience are described elsewhere [201]. A more detailed presentation of group sequential methods for interim analysis of clinical trials may be found in books by Jennison and Turnbull [202] and Proschan, Lan, and Wittes [203].

Curtailed Sampling and Conditional Power Procedures

During the course of monitoring accumulating data, one question often posed is whether the current trend in the data is so impressive that "acceptance" or rejection of

H_0 is already determined, or at least close to being determined. If the results of the trial are such that the conclusions are "known for certain," no matter what the future outcomes might be, then consideration of early termination is in order. A helpful sports analogy is a baseball team "clinching the pennant" after winning a specific game. At that time, it is known for certain who has won and who has not won, regardless of the outcome of the remaining games. Playing the remaining games is done for reasons (e.g., fiscal) other than deciding the winner. This idea has been developed for clinical trials and is often referred to as deterministic curtailed sampling. It should be noted that group sequential methods focus on existing data while curtailed sampling in addition considers the data which have not yet been observed.

Alling [204, 205] may have introduced this concept when he considered the early stopping question and compared the survival experience in two groups. He used the Wilcoxon test for two samples, a frequently used non-parametric test which ranks survival times and which is the basis for one of the primary survival analysis techniques. Alling's method allows stopping decisions to be based on data available during the trial. The trial would be terminated if future data could not change the final conclusion about the null hypothesis. The method is applicable whether all participants are entered at the same time or recruitment occurs over a longer period of time. However, when the average time to the event is short relative to the time needed to enroll participants, the method is of limited value. The repeated testing problem is irrelevant, because any decision is based on what the significance test will be at the end of the study. Therefore, frequent use of this procedure causes no problem with regard to significance level and power.

Many clinical trials with survival time as a response variable have observations that are censored; that is, participants are followed for some length of time and then at some point, no further information about the participant is known or collected. Halperin and Ware [206] extended the method of Alling to the case of censored data, using the Wilcoxon rank statistic. With this method, early termination is particularly likely when the null hypothesis is true or when the expected difference between groups is large. The method is shown to be more effective for small sample sizes than for large studies. The Alling approach to early stopping has also been applied to another commonly used test, the Mantel–Haenszel statistic. However, the Wilcoxon statistic appears to have better early stopping properties than the Mantel–Haenszel statistic.

A deterministic curtailed procedure has been developed [207], for comparing the means of two bounded random variables using the two sample t-test. It assumes that the response must be between two values, A and B ($A<B$). An approximate solution is an extreme case approach. First, all the estimated remaining responses in one group are given the maximum favorable outcome and all the remaining responses in the other take on the worst response. The statistic is then computed. Next, the responses are assigned in the opposite way and a second statistic is computed. If neither of these two extreme results alters the conclusion, no additional data are necessary for testing the hypothesis. While this deterministic curtailed approach provides an answer to an interesting question, the requirement for absolute certainty results in a very conservative test and allows little opportunity for early termination.

In some clinical trials, the final outcome may not be absolutely certain, but almost so. To use the baseball analogy again, a first place team may not have clinched the pennant but is so many games in front of the second place team that it is highly unlikely that it will not, in fact, end up the winner. Another team may be so far behind that it cannot "realistically" catch up. In clinical trials, this idea is often referred to as stochastic curtailed sampling or conditional power. It is identical to the concept of conditional power discussed in the section on extending a trial.

One of the earliest applications of the concept of conditional power was in the CDP [7, 24]. In this trial, several treatment arms for evaluating cholesterol lowering drugs produced negative trends in the interim results. Through simulation, the probability of achieving a positive or beneficial result was calculated given the observed data at the time of the interim analysis. Unconditional power is the probability at the beginning of the trial of achieving a statistically significant result at a prespecified alpha level and with a prespecified alternative treatment effect. Ideally, trials should be designed with a power of 0.80–0.90 or higher. However, once data begin to accumulate, the probability of attaining a significant result increases or decreases with emerging positive or negative trends. Calculating the probability of rejecting the null hypothesis of no effect once some data are available is conditional power.

Lan et al. [208] considered the effect of stochastic curtailed or conditional power procedures on Type I and Type II error rates. If the null hypothesis, H_0, is tested at time t using a statistic, $S(t)$, then at the scheduled end of a trial at time T, the statistic would be $S(T)$. Two cases are considered. First, suppose a trend in favor of rejecting H_0 is observed at time $t<T$, with intervention doing better than control. One then computes the conditional probability, γ_0 of rejecting H_0 at time T; that is, $S(T)>Z_\alpha$, assuming H_0 to be true and given the current data, $S(t)$. If this probability is sufficiently large, one might argue that the favorable trend is not going to disappear. Second, suppose a negative trend or data consistent with the null hypothesis of no difference, at some point t. Then, one computes the conditional probability, γ_1, of rejecting H_0 at the end of the trial, time T, given that some alternative H_1 is true, for a sample of reasonable alternatives. This essentially asks how large the true effect must be before the current "negative" trend is likely to be reversed. If the probability of a trend reversal is highly unlikely for a realistic range of alternative hypotheses, trial termination might be considered.

Because there is a small probability that the results will change, a slightly greater risk of a Type I or Type II error rate will exist than would be if the trial continued to the scheduled end [209]. However, it has been shown that the Type I error is bounded very conservatively by α/γ_0 and the Type II error by β/γ_1. For example, if the probability of rejecting the null hypothesis, given the existing data were 0.85, then the actual Type I error would be no more than 0.05/0.85 or 0.059, instead of 0.05. The actual upper limit is considerably closer to 0.05, but that calculation requires computer simulation. Calculation of these probabilities is relatively straightforward and the details have been described by Lan and Wittes [210]. A summary of these methods, using the approach of DeMets [211], follows:

Let $Z(t)$ represent the standardized statistic at information fraction t. The information fraction may be defined, for example, as the proportion of expected participants

or events observed so far. The conditional power, CP, for some alternative intervention effect θ, using a critical value of $Z\alpha$ for a Type I error of alpha, can be calculated as

$$\mathrm{P}\left[Z(1) \geq Z_{\alpha} \middle| Z(t), \theta\right] = 1 - \Phi\left\{\left|Z_{\alpha} - Z(t)\sqrt{t} - \theta(1-t)\right| / \sqrt{1-t}\right\}$$
$$\text{where } \theta = E(Z(t=1))$$

The alternative θ is defined for various outcomes as follows for:

1. Survival outcome (D = total events)

$$\theta = \sqrt{D/4}\,\mathrm{L\,n}(\lambda_{\mathrm{C}} / \lambda_{T})$$

λ_{C} and λ_{I} are the hazard rates in the control and intervention arms, respectively.

2. Binomial outcome (n = total sample size)

$$\theta = \frac{P_{\mathrm{C}} - P_{T}}{\sqrt{2\overline{p}(1-\overline{p})}/(n/2)} = \frac{(P_{\mathrm{C}} - P_{T})\sqrt{N/4}}{\sqrt{\overline{p}(1-\overline{p})}}$$
$$= 1/2\frac{(P_{\mathrm{C}} - P_{T})\sqrt{N}}{\sqrt{\overline{p}\overline{q}}}$$

where P_{C} and P_{I} are the event rates in the control arm and intervention arm respectively and $\overline{P}$ is the common event rate.

3. Continuous outcome (means) (N = total sample size)

$$\theta = \left(\frac{\mu_{\mathrm{C}} - \mu_{T}}{\sigma}\right)\sqrt{N/4}$$
$$= \frac{1}{2}\left(\frac{\mu_{\mathrm{C}} - \mu_{\mathrm{T}}}{\sigma}\right)\sqrt{N}$$

where μ_{C} and μ_{I} are the mean response levels for the control and the intervention arms, respectively, and σ is the common standard deviation.

If we specify a particular value of the conditional power as γ, then a boundary can also be produced which would indicate that if the test statistic fell below that, the chance of finding a significant result at the end of the trial is less than γ [211]. For example, in Fig. 16.11, the lower futility boundary is based on a specified conditional power γ, ranging from 10 to 30% that might be used to claim futility of finding a positive beneficial claim at the end of the trial. For example, if the standardized statistic crosses that 20% lower boundary, the conditional power for a beneficial result at the end of the trial is less than 0.20 for the specified alternative.

Conditional power calculations are done for a specific alternative but in practice, a monitoring committee would likely consider a range of possibilities. These specified alternatives may range between the null hypothesis of no effect and the

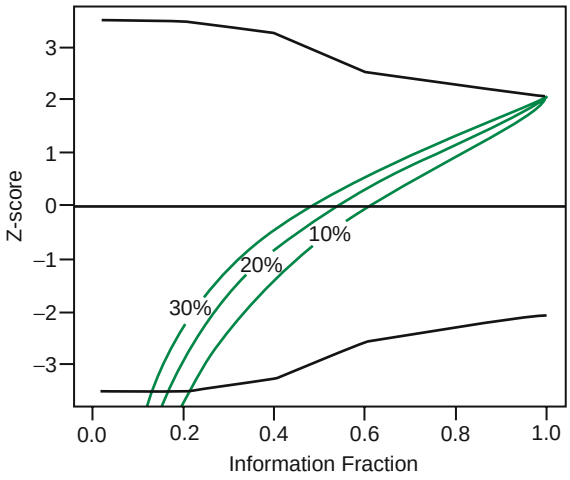


Fig. 16.11 Conditional power boundaries: outer boundaries represent symmetric O'Brien–Fleming type sequential boundaries ($\alpha = 0.05$). Three lower boundaries represent boundaries for 10–30% conditional power to achieve a significant ($p < 0.05$) result of the trial conclusion. [211]

prespecified design based alternative treatment effect. In some cases, a monitoring committee may consider even more extreme beneficial effects to determine just how much more effective the treatment would have to be to raise the conditional power to desired levels. These conditional power results can be summarized in a table or a graph, and then monitoring committee members can assess whether they believe recovery from a substantial negative trend is likely.

Conditional power calculations were utilized in the Vesnarinone in Heart Failure Trial (VEST) [212]. In Table 16.1, the test statistics for the logrank test are provided for the information fractions at a series of monitoring committee meetings. Table 16.2 provides conditional power for VEST at three of the interim analyses. A range of intervention effects was used including the beneficial effect (hazard rate less than 1) seen in a previous vesnarinone trial to the observed negative trend (hazard rates of 1.3

Table 16.1 Accumulating results for the Vesnarinone in Heart Failure Trial (VEST) [212]

Information fraction	Log-rank *Z*-value (high dose)
0.43	+0.99
0.19	−0.25
0.34	−0.23
0.50	−2.04
0.60	−2.32
0.67	−2.50
0.84	−2.22
0.20	−2.43
0.95	−2.71
1.0	−2.41

Table 16.2 Conditional power for Vesnarinone in Heart Failure Trial (VEST) [212]

	Information fraction		
RR	0.50	<0.67	0.84
0.50	0.46	<0.01	<0.01
0.70	0.03	<0.01	<0.01
1.0	<0.01	<0.01	<0.01
1.3	<0.01	<0.01	<0.01
1.5	<0.01	>0.01	<0.01

RR relative risk

and 1.5). It is clear that the conditional power for a beneficial effect was very low by the midpoint of this trial for a null effect or worse. In fact, the conditional power was not encouraging even for the original assumed effect. As described by DeMets et al. [83] the trial continued beyond this point due to the existence of a previous trial that indicated a large reduction in mortality, rather than the harmful effect observed in VEST.

The Beta-Blocker Heart Attack Trial [73, 74] made considerable use of this approach. As discussed, the interim results were impressive with 1 year of follow-up still remaining. One question posed was whether the strong favorable trend ($Z=2.82$) could be lost during that year. The probability of rejecting H_0 at the scheduled end of the trial, given the existing trend (γ_0), was approximately 0.90. This meant that the false positive or Type I error was no more than $\alpha/\gamma_0=0.05/0.90$ or 0.056.

Other Approaches

Other techniques for interim analysis of accumulating data have also received attention. These include binomial sampling strategies [111], decision theoretic models [213], and likelihood or Bayesian methods [214–223]. Bayesian methods require specifying a prior probability on the possible values of the unknown parameter. The experiment is performed and based on the data obtained, the prior probability is adjusted. If the adjustment is large enough, the investigator may change his opinion (i.e., his prior belief). Spiegelhalter et al. [223] and Freedman et al. [217] have implemented Bayesian methods that have frequentist properties very similar to boundaries of either the Pocock or O'Brien–Fleming type. It is somewhat reassuring that two methodologies, even from a different theoretical framework, can provide similar monitoring procedures. While the Bayesian view is critical of the hypothesis testing methods because of the arbitrariness involved, the Bayesian approach is perhaps hampered mostly by the requirement that the investigator formally specify a prior probability. However, if a person during the decision-making process

uses all of the factors and methods discussed in this chapter, a Bayesian approach is involved, although in a very informal way.

One Bayesian method to assess futility that has been used extensively is referred to as predictive power and is related to the concept of conditional power. In this case, the series of possible alternative intervention effects, θ, are represented by a prior distribution for θ, distributing the probability across the alternatives. The prior probability distribution can be modified by the current trend to give an updated posterior for θ. The conditional power is calculated as before for a specific value of θ. Then as shown below, a predictive or "average" power is calculated by integrating the conditional power over the posterior distribution for θ. This predictive power can then be utilized by the monitoring committee to assess whether the trial is still viable.

$$p(X_f \in R) \mid x_0) = \int p\left(X_{\mathrm{f}} \in R \middle| \theta\right) p\left(\theta \middle| x_0\right) \mathrm{d}\theta$$

This predictive power was computed for the various interim analyses conducted in VEST [212] as shown in Table 16.3. In this case, the prior was taken from an earlier trial of vesnarinone where the observed reduction in mortality was over 60% (relative risk = 0.40). For these calculations, the prior was first set at the point estimate of the hazard ratio equal to 0.40. Using this approach, it is clear that VEST would not likely have shown a benefit at the end of the trial.

We have stated that the monitoring committee should be aware of all the relevant information in the use of the intervention which existed before the trial started or which emerges during the course of a trial. Some have argued that all of this information should be pooled or incorporated and updated sequentially in a formal statistical manner [224]. This is referred to as cumulative meta-analysis issues (see Chap. 17). We do not generally support cumulative or sequential meta-analysis as a primary approach for monitoring a trial. We believe that the results of the ongoing trial should be first presented alone in the details described earlier. As supportive evidence for continuation or termination, other analysis may be used, including a pooled analysis of all available external data.

Table 16.3 Predictive probability for the Vesnarinone in Heart Failure Trial (VEST) [212]

		Probability
Date	T*	Hazard rate = 0.40
2/7/96	0.50	0.28
3/7/96	0.60	0.18
4/10/96	0.67	<0.0001
5/19/96	0.84	<0.0001
6/26/96	0.90	<0.0001

* T = Information Fraction

Trend Adaptive Designs and Sample Size Adjustments

Traditionally, sample size adjustments based on comparing emerging trends in the intervention and control groups were discouraged, but statistical methodology has developed that allows trialists to adjust the sample size and maintain the Type I error while regaining power. These methods may be implemented by the monitoring committee or some other third party that is aware of the emerging trend. In general, we do not recommend that the monitoring committee perform this function because it may be aware of other factors that would mitigate any sample size increase but cannot share those issues with the trial investigators or sponsors. This can present an awkward if not an ethical dilemma. Rather, we prefer that a third party who only knows the emerging trend to make the sample size adjustment recommendation to the investigators. Whatever trend adaptive method is used must also take into account the final analyses as discussed briefly in Chap. 17, because it can affect the final critical value.

We will briefly describe several of these methods [101, 102, 225–228].

Using the method proposed by Cui et al. [101], suppose we measuring an outcome variable denoted as X where X has a $N(0,1)$ distribution and n is current sample size, N_0 is initial total sample size, N is new target sample size, θa is hypothesized intervention effect, and t is n/N_0.

In this case, we can have an estimate of the intervention effect and a test statistic based on n observations.

$$\hat{\theta} = \sum_{i}^{n} x_i / n$$

$$z^{(n)} = \sum_{i}^{n} x_i / \sqrt{n}$$

Now, compute a revised sample size N based on the current trend, assuming the same initial Type I error and desired power. A new test statistic is defined that combines the already observed data and the yet to obtained data.

$$Z_{\mathrm{W}}^{(N)} = \sqrt{t} Z^{(n)} + \sqrt{1-t}(N-n)^{-\frac{1}{2}} \sum_{n+1}^{N} x_1$$

In this setting, we would reject the null hypothesis H_0 of no treatment $Z_{\mathrm{W}}^{(n)} > Z_{\alpha}$ effect if this revised test statistic controls the Type I error at the desired level α. However, less weight is assigned to the new or additional observations. This discounting may not be acceptable for scientific and ethical reasons. It is often very challenging to get the design assumptions close enough to what actually happens so that adjustments of this type are not necessary or useful. For example, one observation is that the event is often less than expected, and the intervention effect not as great as assumed. Tsiatis and Mehta [106] have argued that a properly designed group sequential trial is more efficient than these adaptive designs.

A modification proposed by Chen et al. [104] requires that both the weighted and un weighted test statistics exceed the standard critical value.

$$Z^{(N)} \text{ and } Z_{\mathrm{W}}^{(N)} > Z_{\alpha}$$

In this case, the Type I error $< \alpha$ and there is no loss of power.

Another approach, an adjusted p value method, proposed by Proschan and colleagues [102, 228] is to require a "promising" p value before allowing an increase in sample size. However, this approach requires stopping if the first stage p value is not promising. It also requires a larger critical value at the second stage to control the Type I error. As an example, consider a one sided significance level $\alpha = 0.05$. In this case the promising p value, p', and the final critical values are as follows, regardless of the sample size n^2 in the second stage:

p':	0.10	0.15	0.20	0.25	0.50
Z':	1.77	1.82	1.85	1.875	1.95

This simple method will control the Type I error but in fact may make Type I error substantially less than 0.05. A method can be developed to obtain exact Type I error as a function of $Z(t)$ and the adjusted sample size N, using a conditional power type calculation [210] to be described below.

Conditional power is a useful calculation to assess the likelihood of exceeding a critical value at the scheduled end of a trial, given the current data or value of the interim test statistic and making assumptions about the future intervention effect as described earlier in this chapter [162, 208, 210]. The computation of conditional power, CP, in this case is relatively simple. Let $\theta =$ a function of the intervention affect, as described earlier, and then

$$\begin{aligned} \mathrm{CP}(Z(t),\theta) &= P[Z(T) \geq Z_{\alpha} \,|\, Z(t),\theta] \\ &= 1 - \Phi\left\{ Z_{\alpha} - Z(t)\sqrt{t} - \theta(1-t) \right] / \sqrt{(1-t)} \Big\} \end{aligned}$$

Applying the idea of conditional power to the trend adaptive design, we can define an algorithm to adjust the sample size and still control the Type I error [227]. For example,

Let $\Delta =$ observed effect

$\delta =$ assumed effect

If we observe that for $\theta(\Delta)$ as a function of the observed effect Δ, and $\theta(\delta)$ as a function of the assumed δ, then if

$$\begin{aligned} &\mathrm{CP}(Z(t),\theta(\Delta)) > 1.2\mathrm{CP}(Z(t),\theta(\delta)), \quad \text{decrease } N \\ &\mathrm{CP}(Z(t),\theta(\Delta)) < 0.8\mathrm{CP}(Z(t),\theta(\delta)), \quad \text{increase } N \end{aligned}$$

The properties of this procedure have not been well investigated but the idea is related to other conditional power approaches [103]. These conditional power procedures adjust the sample size if the computed conditional power for the current trend is marginal, with only a trivial impact on Type I error. For example, define a lower limit (c_{l}) and an upper limit (c_{u}) such that for the current trend $\theta(\Delta)$:

if $CP(Z(t), \theta(\Delta)) < c_l$, then terminate for futility and accept the null (required),
if $CP(Z(t), \theta(\Delta)) > c_u$, then continue with no change in sample size, or
if $c_l < CP(Z(t), \theta(\Delta)) < c_u$, then increase sample size from N_0 to N to get conditional power to the desired level.

Chen et al. [104] suggested a modest alternative. If the conditional power is 50% or larger, then increase the sample size to get the desired power. An upper cap is typically placed on the size of the increase in sample size. Increase N_0 if the interim result is "promising," defined as conditional power >50% for the current trend but the increase in N_0 cannot be greater than 1.75-fold. Under these conditions, Type I error is not increased and there is no practical loss in power. This approach is one that we favor since it is simple to implement, easy to understand and preserves the design characteristics.

A clear need exists for adaptive designs, including trend adaptive designs. We are fortunate that technical advances have been made through several new methods. Perhaps the largest challenge is how to implement the trend adaptive design without introducing bias or leaving the door open for bias. If one utilizes one of the described trend adaptive designs, anyone who knows the details of the method can "reverse engineer" the implementation and obtain a reasonable estimate of what the current trend ($Z(t)$) must have been to generate the adjusted sample size (N). Given that these trend adaptive designs have as yet not been widely used, there is not enough experience to recommend what can be done to best minimize bias. However, as suggested earlier, a third party who knows only the emerging trend and none of the other secondary or safety data are probably best suited to make these calculations and provide them to the investigators.

References

1. Heart Special Project Committee. Organization, review and administration of cooperative studies (Greenberg Report): A report from the Heart Special Project Committee to the National Advisory Council, May 1967. *Control Clin Trials* 1988;9:137–148.
2. Baum M, Houghton J, Abrams K. Early stopping rules – clinical perspectives and ethical considerations. *Stat Med* 1994;13:1459–1470.
3. Fleming TR, DeMets DL. Monitoring of clinical trials: Issues and recommendations. *Control Clin Trials* 1993;14:183–197.
4. Ellenberg S, Fleming T, DeMets D. *Data Monitoring Committees in Clinical Trials: A Practical Perspective*. West Sussex, England: John Wiley & Sons, Ltd, 2002.
5. DeMets DL, Furberg CD, Friedman L. *Data Monitoring in Clinical Trials: A Case Studies Approach*. New York, NY: Springer Science + Business Media, 2006.
6. Fisher MR, Roecker EB, DeMets DL. The role of an independent statistical analysis center in the industry-modified National Institutes of Health model. *Drug Inf J* 2001;35:115–129.
7. The Coronary Drug Project Research Group. Practical aspects of decision making in clinical trials: The Coronary Drug Project as a case study. *Control Clin Trials* 1981;1:363–376.
8. DeMets DL. Data monitoring and sequential analysis – an academic perspective. *J Acquir Immune Defic Syndr* 1990;3(Suppl 2):S124–S133.
9. Fleming TR, Green SJ, Harrington DP. Considerations for monitoring and evaluating treatment effects in clinical trials. *Control Clin Trials* 1984;5:55–66.

10. Friedman L. The NHLBI model. A 25 year history. *Stat Med* 1993;12:425–431.
11. Geller NL, Stylianou M. Practical issues in the data monitoring of clinical trials: Summary of responses to a questionnaire at NIH. *Stat Med* 1993;12:543–551.
12. George SL. A survey of monitoring practices in cancer clinical trials. *Stat Med* 1993;12: 435–450.
13. O'Neill RT. Some FDA perspectives on data monitoring in clinical trials in drug development. *Stat Med* 1993;12:601–608.
14. Parmar MKB, Machin D. Monitoring clinical trials: Experience of, and proposals under consideration by, the Cancer Therapy Committee of the British Medical Research Council. *Stat Med* 1993;12:497–504.
15. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br J Cancer* 1976;34:585–612.
16. Pocock SJ. Statistical and ethical issues in monitoring clinical trials. *Stat Med* 1993;12:1459–1469.
17. Robinson J. A lay person's perspective on starting and stopping clinical trials. *Stat Med* 1994;13:1473–1477.
18. Rockhold FW, Enas GG. Data monitoring and interim analyses in the pharmaceutical industry: Ethical and logistical considerations. *Stat Med* 1993;12:471–479.
19. Souhami RL. The clinical importance of early stopping of randomized trials in cancer treatments. *Stat Med* 1994;13:1293–1295.
20. Task Force of the Working Group on Arrhythmias of the European Society of Cardiology. The early termination of clinical trials: Causes, consequences, and control. With special reference to trials in the field of arrhythmias and sudden death. *Circulation* 1994;89:2892–2907.
21. Williams GW, Davis RL, Getson AJ, et al. Monitoring of clinical trials and interim analyses from a drug sponsor's point of view. *Stat Med* 1993;12:481–492.
22. Burke G. Discussion of "Early stopping rules – clinical perspectives and ethical considerations." *Stat Med* 1994;13:1471–1472.
23. Buyse M. Interim analyses, stopping rules and data monitoring in clinical trials in Europe. *Stat Med* 1993;12:509–520.
24. Canner PL. Monitoring of the data for evidence of adverse or beneficial treatment effects. *Control Clin Trials* 1983;4:467–483.
25. Simon R. Some practical aspects of the interim monitoring of clinical trials. *Stat Med* 1994;13:1401–1409.
26. Crowley J, Green S, Liu PY, Wolf M. Data monitoring committees and the early stopping guidelines: The Southwest Oncology Group experience. *Stat Med* 1994;13:1391–1399.
27. Green S and Crowley J. Data monitoring committees for Southwest Oncology Group clinical trials. *Stat Med* 1993;12:451–455.
28. Harrington D, Crowley J, George SL, et al. The case against independent monitoring committees. *Stat Med* 1994;13:1411–1414.
29. Herson J. Data monitoring boards in the pharmaceutical industry. *Stat Med* 1993;12:555–561.
30. Pater JL. The use of data monitoring committees in Canadian trial groups. *Stat Med* 1993;12: 505–508.
31. Walters L. Data monitoring committees: The moral case for maximum feasible independence. *Stat Med* 1993;12:575–580.
32. Wittes J. Behind closed doors: The data monitoring board in randomized clinical trials. *Stat Med* 1993;12:419–424.
33. DeMets DL, Fleming TR, Whitley RJ, et al. The Data and Safety Monitoring Board and Acquired Immune Deficiency Syndrome (AIDS) clinical trials. *Control Clin Trials* 1995;16:408–421.
34. Ellenberg SS, Myers MW, Blackwelder WC, Hoth DF. The use of external monitoring committees in clinical trials of the National Institute of Allergy and Infectious Diseases. *Stat Med* 1993;12:461–467.
35. Packer M, Carver JR, Rodeheffer et al. for the PROMISE Study Research Group. Effect of oral milrinone on mortality in severe chronic heart failure. *N Engl J Med* 1991;325:1468–1475.

36. Packer M, Rouleau J, Swedberg K, et al. for the PROFILE Investigators. Effect of Flosequinan on survival in chronic heart failure: Preliminary results of the PROFILE study. *Circulation* 1993; 88(Suppl I): I-301.
37. Packer M, O'Connor CM, Ghali JK, et al. for the Prospective Randomized Amlodipine Survival Evaluation Study Group. Effect of amlodipine on morbidity and mortality in severe chronic heart failure. *N Engl J Med* 1996;335:1107–1114.
38. Shalala D. Protecting research subjects – what must be done. *N Engl J Med* 2000;343:808-810.
39. National Institutes of Health. NIH policy for data and safety monitoring. NIH Guide. http://grants2.nih.gov/grants/guide/notice-files/not98-084.html, 1998.
40. US Food and Drug Administration. Guidance for clinical trial sponsors: Establishment and operation of clinical trial data monitoring committees. http://www.fda.gov/downloads/RegulatoryInformation/Guidances/UCM127073.pdf.
41. Clemens F, Elbourne D, Darbyshire J, Pocock S, the DAMOCLES Group. Data monitoring in randomized controlled trials: Surveys of recent practice and policies. *Clin Trials* 2005;2:22–33.
42. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med* 1987;317:141–145.
43. Meinert CL. Masking monitoring in clinical trials – blind stupidity? *N Engl J Med* 1998;338:1381–1382.
44. Li Z, Geller NL. On the choice of times for data analysis in group sequential clinical trials. *Biometrics* 1991;47:745–750.
45. Bross I. Sequential medical plans. *Biometrics* 1952;8:188–205.
46. Robbins H. Some aspects of sequential design of experiments. *Bull Am Math Soc* 1952;58:527–535.
47. Anscombe FJ. Sequential medical trials. *J Am Stat Assoc* 1963;58:365–383.
48. Armitage P. Restricted sequential procedures. *Biometrika* 1957;44:9–26.
49. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. *J R Stat Soc Ser A* 1969;132:235–244.
50. Robbins H. Statistical methods related to the law or iterated logarithm. *Ann Math Stat* 1970;41:1397–1409.
51. The Coronary Drug Project Research Group. Clofibrate and niacin in coronary heart disease. *JAMA* 1975;231:360–381.
52. Pocock SJ. When to stop a clinical trial. *Br Med J* 1992;305:235–240.
53. DeMets D. Stopping guidelines vs. stopping rules: A practitioner's point of view. *Commun Stat Theory Methods* 1984;13:2395–2417.
54. Pocock SJ. When (not) to stop a clinical trial for benefit. *JAMA* 2005;294:2228–2230.
55. Montori VM, Devereaux PJ, Adhikari NKJ, et al. Randomized trials stopped early for benefit: a systematic review. *JAMA* 2005;294:2203–2209.
56. Freidlin B, Korn EL. Stopping clinical trials for benefit: Impact on estimation. *Clin Trials* 2009;6:119–125.
57. Goodman SN. Stopping early for efficacy: An almost unbiased view. *Clin Trials* 2009;6:133–135.
58. Meinert CL, Knatterud GL, Klimt CR. A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. II. Mortality results. *Diabetes* 1970;19(Suppl):787–830.
59. Knatterud GL, Meinert CL, Klimt CR, et al. Effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. IV. A preliminary report on phenformin results. *JAMA* 1971;217:777–784.
60. Report of the committee for the assessment of biometric aspects of controlled trials of hypoglycemic agents. *JAMA* 1975;231:583–608.
61. Kolata GB. Controversy over study of diabetes drugs continues for nearly a decade. *Science* 1979;203:986–990.
62. The Coronary Drug Project Research Group. The Coronary Drug Project: initial findings leading to modifications of its research protocol. *JAMA* 1970;214:1303–1313.

63. The Coronary Drug Project Research Group. The Coronary Drug Project: Findings leading to further modifications of its protocol with respect to dextrothyroxine. *JAMA* 1972;220: 996–1008.
64. The Coronary Drug Project Research Group. The Coronary Drug Project: Findings leading to discontinuation of the 2.5-mg/day estrogen group. *JAMA* 1973;226:652–657.
65. Pocock SJ, Wang D, Wilhelmsen L, Hennekens CH. The data monitoring experience in the Candasartan Heart failure Assessment of Reduction in Mortality and morbidity (CHARM). *Am Heart J* 2005;149:939–943.
66. Nocturnal Oxygen Therapy Trial Group. Continuous or nocturnal oxygen therapy in hypoxemic chronic obstructive lung disease: a clinical trial. *Ann Intern Med* 1980;93:391–398.
67. DeMets DL, Williams GW, Brown BW Jr, for the NOTT Research Group. A case report of data monitoring experience: The nocturnal oxygen therapy trial. *Control Clin Trials* 1982;3:113–124.
68. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet* 1988;332:349–360.
69. The Diabetic Retinopathy Study Research Group. Diabetic Retinopathy Study. Report no 6. Design, methods, and baseline results. *Invest Ophthalmol Vis Sci* 1981;21:149–209.
70. The Diabetic Retinopathy Study Research Group. Preliminary report on effects of photocoagulation therapy. *Am J Opththalmol* 1976;81:383–396.
71. The Diabetic Retinopathy Study Research Group. Photocoagulation treatment of proliferative diabetic retinopathy: The second report of the Diabetic Retinopathy Study findings. *Ophthalmology* 1978;85:82–106.
72. Ederer F, Podgor MJ, The Diabetic Retinopathy Study Research Group. Assessing possible late treatment effects in stopping a clinical trial early: Diabetic retinopathy study report no. 9. *Control Clin Trials* 1984;5:373–381.
73. Beta-blocker Heart Attack Trial Research Group. A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. *JAMA* 1982;247:1707–1714.
74. DeMets DL, Hardy R, Friedman LM, Lan KKG. Statistical aspects of early termination in the Beta-Blocker Heart Attack Trial. *Control Clin Trials* 1984;5:362–372.
75. Packer M, Coats AJ, Fowler MB, et al., for the Carvedilol Prospective Randomized Cumulative Survival (COPERNICUS) Study Group. Effect of carvedilol on survival in severe chronic heart failure. *N Engl J Med* 2001;334:1651–1658.
76. MERIT-HF Study Group. Effect of metoprolol CR/XL in chronic heart failure: Metoprolol CR/XL Randomised Intervention Trials in congestive heart failure (MERIT-HF). *Lancet* 1999;353:2001–2007.
77. CIBIS-II Investigators and Committees. The Cardiac Insufficiency Bisoprolol study II (CIBIS-II). A randomised trial. *Lancet* 1999;353:9–13.
78. The Beta-Blocker Evaluation Survival Trial Investigators. A trial of the beta-blocker bucindolol in patients with advanced chronic heart failure. *N Engl J Med* 2001;344: 1659–1667.
79. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N Engl J Med* 1989;321:406–412.
80. Pawitan Y, Hallstrom A. Statistical interim monitoring of the Cardiac Arrhythmia Suppression Trial. *Stat Med* 1990;9:1081–1090.
81. Friedman L, Bristow JD, Hallstrom A, et al. Data monitoring in the Cardiac Arrhythmia Suppression Trial. *Online J Curr Clin Trials* 1993 Jul 31;Doc no 79.
82. The Cardiac Arrhythmia Suppression Trial II Investigators. Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. *N Engl J Med* 1992;327:227–233.
83. DeMets DL, Pocock S, Julian DG. The agonizing negative trend in monitoring clinical trials. *Lancet* 1999;354:1983–1988.
84. Swedberg K, Held P, Kjekhus J, et al. Effects of early administration of enalapril on mortality in patients with acute myocardial infarction. Results of the Cooperative New Scandinavian Enalapril Survival Study II (CONSENSUS II). *N Engl J Med* 1992;327:678–684.

85. Furberg C, Campbell R, Pitt B. ACE inhibitors after myocardial infarction (Letter). *N Engl J Med* 1993;328:967–968.
86. Sylvester R, Bartelink H, Rubens R. A reversal of fortune: Practical problems in the monitoring and interpretation of an EORTC breast cancer trial. *Stat Med* 1994;13:1329–1335.
87. Pater JL. Timing the collaborative analysis of three trials comparing 5-FU plus folinic acid (FUFA) to surgery alone in the management of resected colorectal cancer: A National Cancer Institute of Canada Clinical Trials Group (NCIC-CTG) perspective. *Stat Med* 1994;13:1337–1340.
88. Hypertension Detection and Follow-up Program Cooperative Group. Five-year findings of the hypertension detection and follow-up program. Reduction in mortality with high blood pressure, including mild hypertension. *JAMA* 1979;242:2562–2571.
89. Aspirin Myocardial Infarction Study Research Group. A randomized, controlled trial of aspirin in persons recovered from myocardial infarction. *JAMA* 1980;243:661–669.
90. Multiple Risk Factor Intervention Trial Research Group. Multiple risk factor interventional trial. Risk factor changes and mortality results. *JAMA* 1982;248:1465–1477.
91. CASS Principal Investigators and their Associates. Coronary Artery Surgery Study (CASS): A randomized trial of coronary artery bypass surgery. Survival data. *Circulation* 1983;68:939–950.
92. Multiple Risk Factor Intervention Trial Research Group. Mortality after 16 years for participants randomized to the Multiple Risk Factor Intervention Trial. *Circulation* 1996;94: 946–951; correction 1997;95:760.
93. Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the Women's Health Initiative randomized clinical trial. *JAMA* 2002;288:321–333.
94. The Women's Health Initiative Steering Committee. Effects of conjugated equine estrogen in postmenopausal women with hysterectomy. The Women's Health Initiative randomized controlled trial. *JAMA* 2004;291:1701–1712.
95. McMurray JJ, Teerlink JR, Cotter G, for the VERITAS Investigators. Effects of tezosentan on symptoms and clinical outcomes in patients with acute heart failure. The VERITAS randomized controlled trials. *JAMA* 2007;298:2009–2019.
96. Tegler CH, Furberg CD. Lessons from warfarin trials in atrial fibrillation: Missing the window of opportunity. In DeMets DL, Friedman L, Furberg CD (eds.) *Data Monitoring in Clinical Trials: A Case Studies Approach*. New York: Springer Science Business Media, 2006, pp. 312–319.
97. Liberati A. Conclusions. 1: The relationship between clinical trials and clinical practice: The risks of underestimating its complexity. *Stat Med* 1994;13:1485–1491.
98. O'Neill RT. Conclusions. 2: The relationship between clinical trials and clinical practice: The risks of underestimating its complexity. *Stat Med* 1994;13:1493–1499.
99. Collaborative Group on Antenatal Steroid Therapy. Effect of antenatal dexamethasone administration on the prevention of respiratory distress syndrome. *Am J Obstet Gynecol* 1981;141:276–287.
100. The MIAMI Trial Research Group. Metoprolol In Acute Myocardial Infarction (MIAMI). A randomized placebo-controlled international trial. *Eur Heart J* 1985;6:199–226.
101. Cui L, Hung HM, Wang SJ. Modification of sample size in group sequential clinical trials. *Biometrics* 1999;55:853–857.
102. Proschan MA, Liu Q, Hunsberger S. Practical midcourse sample size modification in clinical trials. *Control Clin Trials* 2003;24:4–15.
103. Lan KKG, Trost DC. Estimation of parameters and sample size re-estimation. In *ASA Proceedings of the Biopharmaceutical Section,* pp. 48–51. American Statistical Association (Alexandria, VA), 1997.
104. Chen JYH, DeMets DL, Lan KKG. Increasing the sample size when the unblinded interim result is promising. *Stat Med* 2004;23:1023–1038.
105. Fleming TR. Standard versus adaptive monitoring procedures: A commentary. *Stat Med* 2006;25:3305–3312; discussion 3313–3314, 3326–3347.

106. Tsiatis AA, Mehta C. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* 2003;90:367–378.
107. Taylor AL, Ziesche S, Yancy C, et al. for the African-American Heart Failure Trial Investigators. Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. *N Engl J Med* 2004;351:2049–2057.
108. Wald A. *Sequential Analysis*. New York: John Wiley and Sons, 1947.
109. Cornfield J. Sequential trials, sequential analysis and the likelihood principle. *Am Stat* 1966;20:18–23.
110. Armitage P. *Sequential Medical Trials* (2nd edition). New York: John Wiley and Sons, 1975.
111. Simon R, Weiss GH, Hoel DG. Sequential analysis of binomial clinical trials. *Biometrika* 1975;62:195–200.
112. Whitehead J, Jones D. The analysis of sequential clinical trials. *Biometrika* 1979;66: 443–452.
113. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. New York: Haisted Press, 1983.
114. Whitehead J, Stratton I. Group sequential clinical trials with triangular continuation regions. *Biometrics* 1983;39:227–236.
115. DeMets DL, Lan KKG. An overview of sequential methods and their application in clinical trials. *Commun Stat Theory Methods* 1984;13:2315–2338.
116. Silverman WA, Agate FJ Jr, Fertig JW. A sequential trial of the nonthermal effect of atmospheric humidity on survival of newborn infants of low birth weight. *Pediatrics* 1963;31:719–724.
117. Truelove SC, Watkinson G, Draper G. Comparison of corticosteroid and sulphasalazine therapy in ulcerative colitis. *Br Med J* 1962;2:1708–1711.
118. Acute Leukemia Group B, Freireich EJ, Gehan E, Frei E, et al. The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: A model for evaluation of other potentially useful therapy. *Blood* 1963;21:699–716.
119. McPherson CK, Armitage P. Repeated significance tests on accumulating data when the null hypothesis is not true. *J R Stat Soc Ser A* 1971;134:15–25.
120. Whitehead J, Jones DR, Ellis SH. The analysis of a sequential clinical trial for the comparison of two lung cancer treatments. *Stat Med* 1983;2:183–190.
121. Dambrosia JM, Greenhouse SW. Early stopping for sequential restricted tests of binomial distributions. *Biometrics* 1983;39:695–710.
122. Chatterjee SK, Sen PK. Nonparametric testing under progressive censoring. *Calcutta Stat Assoc Bull* 1973;22:13–50.
123. Muenz LR, Green SB, Byar DP. Applications of the Mantel–Haenszel statistic to the comparison of survival distributions. *Biometrics* 1977;33:617–626.
124. Davis CE. A two sample Wilcoxon test for progressively censored data. *Commun Stat Theory Methods* 1978;A7:389–398.
125. Koziol J, Petkau J. Sequential testing of the equality of two survival distributions using the modified Savage statistic. *Biometrika* 1978;65:615–623.
126. Breslow N, Haug C. Sequential comparison of exponential survival curves. *J Am Stat Assoc* 1972;67:691–697.
127. Canner PL. Monitoring treatment differences in long-term clinical trials. *Biometrics* 1977;33:603–615.
128. Jones D, Whitehead J. Sequential forms of the log rank and modified Wilcoxon tests for censored data. *Biometrika* 1979;66:105–113.
129. Joe H, Koziol J, Petkau JA. Comparison of procedures for testing the equality of survival distributions. *Biometrics* 1981;37:327–340.
130. Nagelkerke NJD, Hart AAM. The sequential comparison of survival curves. *Biometrika* 1980;67:247–249.
131. Sellke T, Siegmund D. Sequential analysis of the proportional hazards model. *Biometrika* 1983;70:315–326.

132. Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. *Br J Radiol* 1971;44:793–797.
133. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977;64:191–199.
134. Pocock SJ. Size of cancer clinical trials and stopping rules. *Br J Cancer* 1978;38:757–766.
135. Pocock SJ. Interim analyses for randomized clinical trials: The group sequential approach. *Biometrics* 1982;38:153–162.
136. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979;35:549–556.
137. Freedman LS, Lowe D, Macaskill P. Stopping rules for clinical trials. *Stat Med* 1983;2:167–174.
138. DeMets DL. Practical aspects in data monitoring: A brief review. *Stat Med* 1987;6:753–760.
139. Emerson SS, Fleming TR. Interim analyses in clinical trials. *Oncology* 1990;4:126-133.
140. Fleming TR, Watelet LF. Approaches to monitoring clinical trial. *J Natl Cancer Inst* 1989;81:188–193.
141. Jennison C, Turnbull BW. Statistical approaches to interim monitoring of medical trials: A review and commentary. *Stat Sci* 1990;5:299–317.
142. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* 1966;50:163–170.
143. Tsiatis AA. The asymptotic joint distribution of the efficient scores tests for the proportional hazards model calculated over time. *Biometrika* 1981;68:311–315.
144. Tsiatis AA. Repeated significance testing for a general class of statistics used in censored survival analysis. *J Am Stat Assoc* 1982;77:855–861.
145. Tsiatis AA. Group sequential methods for survival analysis with staggered entry. In Johnson R, Crowley J (eds.). *Survival Analysis. Monograph Series 2.* Hayward, California: IMS Lecture Notes, 1982, pp. 257–268.
146. Gail MH, DeMets DL, Slud EV. Simulation studies on increments of the two-sample log rank score test for survival data, with application to group sequential boundaries. In Johnson R, Crowley J (eds.). *Survival Analysis. Monograph Series 2.* Hayward, California: IMS Lecture Notes, 1982, pp. 287–301.
147. Harrington DP, Fleming TR, Green SJ. Procedures for serial testing in censored survival data. In Johnson R, Crowley J (eds.). *Survival Analysis. Monograph Series 2.* Hayward, California: IMS Lecture Notes, 1982, pp. 269–286.
148. Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 1965;52:203–223.
149. Slud E, Wei LJ. Two-sample repeated significance tests based on the modified Wilcoxon statistic. *J Am Stat Assoc* 1982;77:862–868.
150. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *J R Stat Soc Ser A* 1972;135:185–206.
151. DeMets DL, Gail MH. Use of logrank tests and group sequential methods at fixed calendar times. *Biometrics* 1985;41:1039–1044.
152. George SL. Discussion of "Sequential methods based on the boundaries approach for the clinical comparison of survival times." *Stat Med* 1994;13:1369–1370.
153. Kim K. Study duration for group sequential clinical trials with censored survival data adjusting for stratification. *Stat Med* 1992;11:1477–1488.
154. Kim K, Tsiatis AA. Study duration for clinical trials with survival response and early stopping rule. *Biometrics* 1990;46:81–92.
155. Whitehead J. Sequential methods based on the boundaries approach for the clinical comparison of survival times. *Stat Med* 1994;13:1357–1368.
156. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983;70:659–663.
157. Lan KKG, DeMets DL, Halperin M. More flexible sequential and non-sequential designs in long-term clinical trials. *Commun Stat Theory Methods* 1984;13:2339–2354.

158. DeMets DL, Lan KKG. Interim analysis: The alpha spending function approach. *Stat Med* 1994;13:1341–1352.
159. Kim K, DeMets DL. Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* 1987;74:149–154.
160. Lan KKG, DeMets DL. Group sequential procedures: Calendar versus information time. *Stat Med* 1989;8:1191–1198.
161. Lan KKG, Rosenberger WF, Lachin JM. Use of spending functions for occasional or continuous monitoring of data in clinical trials. *Stat Med* 1993;12:2214–2231.
162. Lan KKG, Zucker D. Sequential monitoring of clinical trials: The role of information in Brownian motion. *Stat Med* 1993;12:753–765.
163. Lan KKG, Reboussin DM, DeMets DL. Information and information fractions for design and sequential monitoring of clinical trials. *Commun Stat Theory Methods* 1994;23:403–420.
164. Reboussin DM, DeMets DL, Kim K, Lan KKG. Programs for computing group sequential bounds using the Lan–DeMets method. *Control Clin Trials* 2000;21:190–207.
165. Hwang IK, Shih WJ, De Cani JS. Group sequential designs using a family of type I error probability spending function. *Stat Med* 1990;9:1439–1445.
166. Wang SK, Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 1987;43:193–199.
167. Lan KKG, DeMets DL. Changing frequency of interim analyses in sequential monitoring. *Biometrics* 1989;45:1017–1020.
168. Proschan MA, Follman DA, Waclawiw MA. Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics* 1992;48:1131–1143.
169. Geller NL. Discussion of "Interim analysis: The alpha spending approach." *Stat Med* 1994;13:1353–1356.
170. Falissard B, Lellouch J. A new procedure for group sequential analysis in clinical trials. *Biometrics* 1992;48:373–388.
171. Lan KKG, Lachin J. Implementation of group sequential logrank tests in a maximum duration trial. *Biometrics* 1990;46:759–770.
172. Kim K, DeMets DL. Sample size determination for group sequential clinical trials with immediate response. *Stat Med* 1992;11:1391–1399.
173. Lee JW. Group sequential testing in clinical trials with multivariate observations: a review. *Stat Med* 1994;13:101–111.
174. Lee JW, DeMets DL. Sequential comparison of change with repeated measurement data. *J Am Stat Assoc* 1991;86:757–762.
175. Lee JW, DeMets DL. Sequential rank tests with repeated measurements in clinical trials. *J Am Stat Assoc* 1992;87:136–142.
176. Su JQ, Lachin JM. Group sequential distribution-free methods for the analysis of multivariate observations. *Biometrics* 1992;48:1033–1042.
177. Wei LJ, Su JQ, Lachin JM. Interim analyses with repeated measurements in a sequential clinical trial. *Biometrika* 1990;77:359–364.
178. Wu MC, Lan KKG. Sequential monitoring for comparison of changes in a response variable in clinical studies. *Biometrics* 1992;48:765–779.
179. Gange SJ, DeMets DL. Sequential monitoring of clinical trials with correlated responses. *Biometrika* 1996;83:157–167.
180. Fairbanks K, Madsen R. *P* values for tests using a repeated significance test design. *Biometrika* 1982;69:69–74.
181. Tsiatis AA, Rosner GL, Mehta CR. Exact confidence intervals following a group sequential test. *Biometrics* 1984;40:797–803.
182. Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. *Biometrika* 1990;77:875–892.
183. Hughes MD, Pocock SJ. Stopping rules and estimation problems in clinical trials. *Stat Med* 1988;7:1231–1242.
184. Kim K. Point estimation following group sequential tests. *Biometrics* 1989;45:613–617.

185. Kim K, DeMets DL. Confidence intervals following group sequential tests in clinical trials. *Biometrics* 1987;4:857–864.
186. Pocock SJ, Hughes MD. Practical problems in interim analyses, with particular regard to estimation. *Control Clin Trials* 1989;10(Suppl):209S–221S.
187. Rosner GL, Tsiatis AA. Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika* 1988;75:723–729.
188. Siegmund D. Estimation following sequential tests. *Biometrika* 1978;65:341–349.
189. Whitehead J, Facey KM. Analysis after a sequential trial: A comparison of orderings of the sample space. Presented at the Joint Society for Clinical Trials/International Society for Clinical Biostatistics, Brussels, 1991.
190. Chang MN, O'Brien PC. Confidence intervals following group sequential tests. *Control Clin Trials* 1986;7:18–26.
191. Whitehead J. On the bias of maximum likelihood estimation following a sequential test. *Biometrika* 1986;73:573–581.
192. Jennison C, Turnbull BW. Repeated confidence intervals for group sequential clinical trials. *Control Clin Trials* 1984;5:33–45.
193. Jennison C, Turnbull BW. Interim analyses: The repeated confidence interval approach. *J R Stat Soc Series B Stat Methodol* 1989;51:305–334; discussion: 334–361.
194. DeMets DL, Lan KKG. Discussion of: Interim analyses: The repeated confidence interval approach by C. Jennison and B.W. Turnbull. *J R Stat Soc Series B Stat Methodol* 1989;51:344.
195. Fleming TR. Evaluation of active control trials in AIDS. *J Acquir Immune Defic Syndr* 1990;3(Suppl):S82–S87.
196. Fleming TR. Treatment evaluation in active control studies. *Cancer Treat Rep* 1987;17:1061–1065.
197. DeMets DL, Ware JH. Group sequential methods in clinical trials with a one-sided hypothesis. *Biometrika* 1980;67:651–660.
198. DeMets DL, Ware JH. Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika* 1982;69:661–663.
199. Emerson SS, Fleming TR. Symmetric group sequential test designs. *Biometrics* 1989;45: 905–923.
200. Gould AL, Pecore VJ. Group sequential methods for clinical trials allowing early acceptance of Ho and incorporating costs. *Biometrika* 1982;69:75–80.
201. Feyzi J, Julian D, Wikstrand J, Wedel H. Data monitoring experience in the Metoprolol CR/XL randomized intervention trial in chronic heart failure: Potentially high-risk treatment in high-risk patients. In DeMets DL, Friedman L, Furberg CD (eds.). *Data Monitoring in Clinical Trials: A Case Studies Approach*. New York, NY: Springer Science + Business Media, 2005, pp. 136–147
202. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials.* Boca Raton, FL: Chapman and Hall/CRC, 2000.
203. Proschan MA, Lan KKG, Wittes JT. *Statistical Monitoring of Clinical Trials: A Unified Approach*. New York, NY: Springer Science + Business Media, LLC, 2006.
204. Alling DR. Early decision in the Wilcoxon two sample test. *J Am Stat Assoc* 1963;58:713–720.
205. Alling DW. Closed sequential tests for binomial probabilities. *Biometrika* 1966;53:73–84.
206. Halperin M, Ware J. Early decision in a censored Wilcoxon two-sample test for accumulating survival data. *J Am Stat Assoc* 1974;69:414–422.
207. DeMets DL, Halperin M. Early stopping in the two-sample problem for bounded random variables. *Control Clin Trials* 1982;3:1–11.
208. Lan KKG, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical trials. *Commun Stat. Sequential Anal* 1982;1:207–219.
209. Halperin M, Lan KKG, Ware JH, et al. An aid to data monitoring in long-term clinical trials. *Control Clin Trials* 1982;3:311–323.
210. Lan KKG, Wittes J. The B-value: A tool for monitoring data. *Biometrics* 1988;44:579–585.
211. DeMets DL. Futility approaches to interim monitoring by data monitoring committees. *Clin Trials* 2006;3:522–529.

212. Cohn JN, Goldstein SO, Greenberg BH, et al., for the Vesnarinone Trial Investigators. A dose-dependent increase in mortality with vesnarinone among patients with severe heart failure. *N Engl J Med* 1998;339:1810–1816.
213. Colton T. A model for selecting one of two medical treatments. *J Am Stat Assoc* 1963;58: 388–400.
214. Cornfield J. A Bayesian test of some classical hypotheses – with applications to sequential clinical trials. *J Am Stat Assoc* 1966;61:577–594.
215. Cornfield J. Recent methodological contributions to clinical trials. *Am J Epidemiol* 1976;104:408–421.
216. Choi SC, Pepple PA. Monitoring clinical trials based on predictive probability of significance. *Biometrics* 1989;45:317–323.
217. Freedman LS, Spiegelhalter DJ, Parmar MKB. The what, why, and how of Bayesian clinical trials monitoring. *Stat Med* 1994;13:1371–1383.
218. Grieve AP. Predictive probability in clinical trials. *Biometrics* 1991;47:323–330.
219. George SL, Li C, Berry DA, Green MR. Stopping a clinical trial early: Frequentist and Bayesian approaches applied to a CALGB trial of non-small-cell lung cancer. *Stat Med* 1994;13:1313–1327.
220. Carlin BP, Louis TA. *Bayes and Empirical Bayes Methods for Data Analysis* (2nd edition). Boca Raton, FL: Chapman and Hall/CRC, 2000.
221. Machin D. Discussion of "The what, why and how of Bayesian clinical trials monitoring." *Stat Med* 1994;13:1385–1389.
222. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Stat Med* 1986;5:421–433.
223. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: Conditional or predictive power? *Control Clin Trials* 1986;7:8–17.
224. Lau J, Antman EM, Jimenez-Silva J, et al. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N Engl J Med* 1992;327:248–254.
225. Fisher LD. Self-designing clinical trials. *Stat Med* 1998;17:1551–1562.
226. Shen Y, Fisher L. Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics* 1999;55:190–197.
227. Cui L, Hun HMJ, Wang SJ. Impact of changing sample size in a group sequential clinical trial. Proceedings of the Biopharmaceutical Section, American Statistical Association, 1997, pp. 52–57.
228. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995;51:1315–1324.

Chapter 17
Issues in Data Analysis

The analysis of data obtained from a clinical trial represents the outcome of the planning and implementation already described. Primary and secondary questions addressed by the clinical trial can be tested and new hypotheses generated. Data analysis is sometimes viewed as simple and straightforward, requiring little time, effort, or expense. However, careful analysis usually requires a major investment in all three. It must be done with as much care and concern as any of the design or data-gathering aspects. Furthermore, inappropriate statistical analyses can introduce bias, result in misleading conclusions, and impair the credibility of the trial.

Several introductory textbooks of statistics [1–8] provide excellent descriptions for many basic methods of analysis. Chapter 15 presents essentials for analysis of survival data, since they are frequently of interest in clinical trials and are not covered in most introductory statistics texts. This chapter focuses on some issues in the analysis of data, which seem to cause confusion in the medical research community. Some of the proposed solutions are straightforward; others require judgment. They reflect a point of view developed by the authors and many colleagues in numerous collaborative efforts over three to four decades. Some [9–12] have taken similar positions, whereas others [13, 14] have opposing views on several issues.

The analytic approaches discussed here primarily apply to late phase (III and IV) trials. Various exploratory analysis approaches may be entirely reasonable in early phase (I and II) studies where the goal is to obtain information and insight to design better subsequent trials. However, some the fundamentals presented may still be of value in these early phase trials. We have used early examples that were instrumental in establishing many of the analytic principles and added new examples which reinforce them. However, given the multitude of clinical trials, it is not possible to include all examples.

Fundamental Point

Excluding randomized participants or observed outcomes from analysis and subgrouping on the basis of outcome or other response variables can lead to biased results. Those biases can be of unknown magnitude or direction.

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_17, © Springer Science+Business Media, LLC 2010

Which Participants Should Be Analyzed?

The issue of which participants are to be included in the data analysis often arises in clinical trials. Although a laboratory study may have carefully regulated experimental conditions, even the best designed and managed clinical trial cannot be perfectly implemented. Response variable data may be missing, the protocol may not be completely adhered to, and some participants, in retrospect, will not have met the entrance criteria. Some investigators may, after a trial has been completed, prefer to remove from the analysis participants who did not fit the eligibility criteria or did not follow the protocol perfectly. Conversely, others believe that once a participant is randomized, that participant should always be followed and included in the analysis.

The *intention-to-treat* principle states that all participants randomized and all events, as defined in the protocol, should be accounted for in the primary analysis [9]. This requirement is stated in the International Conference on Harmonisation and FDA guidelines [15, 16]. There are often proposed "modified intention-to-treat" analyses, or "per protocol" or "on treatment" analyses, that suggest that only participants who received at least some of the intervention should be included. However, as we will discuss, any deviations from pure intention-to-treat offer the potential for bias and should be avoided, or at a minimum presented along with a strict intention-to-treat analysis. Many publications claim that the analyses are intention-to-treat when in reality are not. Although the phrase is widely used, "per protocol" analysis suggests that the analysis is the one preferred in the trial's protocol. We think that "on treatment" analysis more accurately reflects what is done.

The rationales for each of these positions are presented in the following pages. This chapter has adopted, in part, the terminology used by Peto and colleagues [9] to classify participants according to the nature and extent of their participation.

Exclusions refer to people who are screened as potential participants for a randomized trial but who do not meet all of the entry criteria and, therefore, are not randomized. Reasons for exclusion might be related to age, severity of disease, refusal to participate, or any of numerous other determinants evaluated before randomization. Since these potential participants are not randomized, their exclusion does not bias any intervention-control group comparison (sometimes called *internal validity*). Exclusions do, however, influence interpretation and applicability of the results of the clinical trial (*external validity*). In some circumstances, follow-up of excluded people can be helpful in determining to what extent the results can be generalized. If the event rate in the control group is considerably lower than anticipated, an investigator may want to determine whether most high risk people were excluded or whether she was incorrect in her initial assumption.

Withdrawals from analysis refer to participants who have been randomized but are deliberately excluded from the analysis. As the fundamental point states, omitting participants from analyses can bias the results of the study [17]. If participants are withdrawn, the burden rests with the investigator to convince the scientific community that the analysis has not been biased. However, this can be a difficult task because

no one can be sure that participants were not differentially withdrawn from the study groups. Differential withdrawal can occur even if the number of omitted participants is the same in each group, since the reasons for withdrawal in each group may be different and thus their risk of primary, secondary, and adverse events. As a result, the participants remaining in the trial may not be comparable, undermining one of the reasons for randomization.

Many reasons are given for withdrawing participants from the analysis such as ineligibility and nonadherence.

Ineligibility

A previously commonly cited reason for withdrawal is that some participants did not meet the entry criteria, a protocol violation unknown at the time of enrollment. Admitting unqualified participants may be the result of a simple clerical error, a laboratory error, a misinterpretation, or a misclassification. Clerical mistakes such as listing wrong sex or age may be obvious. Other errors can arise from differing interpretation of diagnostic studies such as electrocardiograms, X-rays, or biopsies. It is not difficult to find examples in earlier literature [17–26]. This reason for withdrawal used to be common, but appears to be less frequent now, at least in papers published in major journals.

Withdrawals for ineligibility can involve a relatively large number of participants. In an early trial by the Canadian Cooperative Study Group [19], 64 of the 649 enrolled participants (10%) with stroke were later found to have been ineligible. In this four-armed study, the numbers of ineligible participants in the study groups ranged from 10 to 25. The reasons for the ineligibility of these 64 participants were not reported, nor were their outcome experiences. Before cancer cooperative groups implemented phone-in or electronic eligibility checks, 10–20% of participants entered into a trial may have been ineligible after further review. By taking more careful care at the time of randomization, the number of ineligible participants was reduced to a very small percent [27]. Currently, web based systems or Interactive Voice Recording Systems are used for multicenter and multinational clinical trials. These interactive systems can double-check key eligibility criteria before randomization is assigned, cutting down on the ineligibility rate. For example, several trials in chronic heart failure employed these methods [28–31].

A study design may require enrollment within a defined time period following a qualifying event. Because of this time constraint, data concerning a participant's eligibility might not be available or confirmable at the time the decision must be made to enroll him. For example, the Beta-Blocker Heart Attack Trial (BHAT) looked at 2-year mortality in people who were administered a beta-blocking drug during hospitalization for an acute myocardial infarction [20]. Because of known variability in interpretation, the protocol required that the diagnostic electrocardiograms be read by a central unit. However, this verification took several weeks to accomplish. Local institutions, therefore, interpreted the electrocardiograms and

decided whether the patient met the necessary criteria for inclusion. Almost 10% of the enrolled participants did not have their myocardial infarction confirmed according to a central reading, and were "incorrectly" randomized. The question then arose: Should the participants be kept in the trial and included in the analysis of the response variable data? The BHAT protocol required follow-up and analysis of all randomized participants. In this case, the observed benefits from the intervention were similar in those eligible as well as in those "ineligible."

A more complicated situation occurs when the data needed for enrollment cannot be obtained until hours or days have passed, yet the study design requires initiation of intervention before then. For instance, in the Multicenter Investigation of the Limitation of Infarct Size (MILIS) [22], propranolol, hyaluronidase, or placebo was administered shortly after participants were admitted to the hospital with possible acute myocardial infarctions. In some, the diagnosis of myocardial infarction was not confirmed until after electrocardiographic and serum enzyme changes had been monitored for several days. Such participants were, therefore, randomized on the basis of a preliminary diagnosis of infarction. Subsequent testing may not have supported the initial diagnosis. Another example of this problem involves a study of pregnant women who were likely to deliver prematurely and therefore, would have children who were at a higher than usual risk of being born with respiratory distress syndrome [23]. Corticosteroids administered to the mother prior to delivery were hypothesized to protect the premature child from developing this syndrome. Although, at the time of the mother's randomization to either intervention or control groups, the investigator could not be sure that the delivery would be premature, she needed to make a decision whether to enroll the mother into the study. Other examples include trials where thrombolytic agents are being evaluated in reducing mortality and morbidity during and after a myocardial infarction. In these trials, agents must be given rapidly before diagnosis can be confirmed [32].

To complicate matters still further, the intervention given to a participant can affect or change the entry diagnosis. For example, in the above mentioned study to limit infarct size, some participants without a myocardial infarction were randomized because of the need to begin intervention before the diagnosis was confirmed. Moreover, if the interventions succeeded in limiting infarct size, they could have affected the electrocardiogram and serum enzyme levels. Participants in the intervention groups with a small myocardial infarction may have had the infarct size reduced or limited and therefore appeared not to have had an infarction. Thus, they would not seem to have met the entry criteria. However, this situation could not exist in the placebo control group. If the investigators had withdrawn participants in retrospect who did not meet the study criteria for a myocardial infarction, they would have withdrawn more participants from the intervention groups (those with no documented infarction plus those with small infarction) than from the control group (those with no infarction). This would have produced a bias in later comparisons. On the other hand, it could be assumed that a similar number of truly ineligible participants were randomized to the intervention groups and to the control group. In order to maintain comparability, the investigators might have decided to withdraw the same number of participants from each group. The ineligible participants

in the control group could have been readily identified. However, the participants in the intervention groups who were truly ineligible had to be distinguished from those made to appear ineligible by the effects of the interventions. This would have been difficult, if not impossible. In the MILIS for example, all randomized participants were retained in the analysis [22].

An example of possible bias because of withdrawal of ineligible participants is found in the Anturane Reinfarction Trial, which compared sulfinpyrazone with placebo in participants who had recently suffered a myocardial infarction [24–26]. As seen in Table 17.1, of 1,629 randomized participants (813 to sulfinpyrazone, 816 to placebo), 71 were subsequently found to be ineligible. Thirty-eight had been assigned to sulfinpyrazone and 33 to placebo. Despite relatively clear definitions of eligibility and comparable numbers of participants withdrawn, mortality among these ineligible participants was 26.3% in the sulfinpyrazone group (10 of 38) and 12.1% in the placebo group (4 of 33) [25]. The eligible placebo group participants had a mortality of 10.9%, similar to the 12.1% seen among the ineligible participants. In contrast, the eligible participants on sulfinpyrazone had a mortality of 8.3%, less than one-third that of the ineligible participants. Including all 1,629 participants in the analysis gave 9.1% mortality in the sulfinpyrazone group, and 10.9% mortality in the placebo group ($p=0.20$). Withdrawing the 71 ineligible participants (and 14 deaths, 10 vs. 4) gave an almost significant $p=0.07$.

Stimulated by criticisms of the study, the investigators initiated a reevaluation of the Anturane Reinfarction Trial results. An independent group of reviewers examined all reports of deaths in the trial [26]. Instead of 14 deceased participants who were ineligible, it found 19; 12 in the sulfinpyrazone group and seven in the placebo group. Thus, supposedly clear criteria for ineligibility can be judged differently. This trial was an early example that affirmed the value of the intention-to-treat principle.

Three trial design policies that relate to withdrawals because of entry criteria violations have been discussed by Peto et al. [9]. The first policy is not to enroll participants until all the diagnostic tests have been confirmed and all the entry criteria have been carefully checked. Once enrollment takes place, no withdrawals are allowed. For some studies, such as the one on limiting infarct size, this policy cannot be applied because firm diagnoses cannot be ascertained prior to the time when intervention has to be initiated.

The second policy is to enroll marginal or unconfirmed cases and later withdraw those participants who are proven to have been misdiagnosed. This would be allowed, however, only if the decision to withdraw is based on data collected before enrollment.

Table 17.1 Mortality by study group and eligibility status in the Anturane Reinfarction Trial

	Randomized	Percent mortality	Ineligible	Percent mortality	Eligible	Percent mortality
Sulfinpyrazone	813	9.1	38	26.3	775	8.3
Placebo	816	10.9	33	12.1	783	10.9

Any process of deciding upon withdrawal of a participant from a study group should be done blinded with respect to the participant's outcome and group assignment.

A third policy is to enroll some participants with unconfirmed diagnoses and to allow no withdrawals. This procedure is always valid in that the investigator compares two randomized groups which are comparable at baseline. However, this policy is conservative because each group contains some participants who might not benefit from the intervention. Thus, the overall trial may have less power to detect differences of interest.

A modification to these three policies is recommended. Every effort should be made to establish the eligibility of participants prior to any randomization. No withdrawals should be allowed, and the analyses should include all participants enrolled. Analyses based on only those truly eligible may be performed. If the analyses of data from all enrolled participants and from those eligible agree, the interpretation of the results is clear, at least with respect to participant eligibility. If the results differ, however, the investigator must be very cautious in her interpretation. In general, she should emphasize the analysis with all the enrolled participants because that analysis is always valid.

Any policy on withdrawals should be stated in the study protocol before the start of the study. The actual decision to withdraw specific participants should be done without knowledge of the study group, ideally by someone not directly involved in the trial. Of special concern is withdrawal based on review of selected cases, particularly if the decision rests on a subjective interpretation. Even in double-blind trials, blinding may not be perfect, and the investigator may supply information for the eligibility review differentially depending upon study group and health status. Therefore, withdrawal should be done early in the course of follow-up, before a response variable has occurred, and with a minimum of data exchange between the investigator and the person making the decision to withdraw the participant. This withdrawal approach does not preclude a later challenge by readers of the report, on the basis of potential bias. It should, however, remove the concern that the withdrawal policy was dependent on the outcome of the trial. The withdrawal rules should not be based on knowledge of study outcomes. Even when these guidelines are followed, if the number of entry criteria violations is substantially different in the study groups, or if the event rates in the withdrawn participants are different between the groups, the question will certainly be raised whether bias played a role in the decision to withdraw participants.

Nonadherence

Nonadherence to the prescribed intervention or control regimen is another reason that participants are withdrawn from analysis [33–52]. One version of this is to define an *on treatment* analysis that eliminates any participant who does not adhere to the intervention by some specified amount, as defined in the protocol. One form of nonadherence may be drop-outs and drop-ins (Chap. 14). Drop-outs are participants in

the intervention arm who do not adhere to the regimen. Drop-ins are participants in the control arm who start to use the intervention. The decision not to adhere to the protocol intervention may be made by the participant, his primary care physician, or the trial investigator. Nonadherence may be due to adverse events in either the intervention or control group, loss of participant interest or perceived benefit, changes in the underlying condition of a participant, or a variety of other reasons.

Withdrawal from analysis of participants who do not adhere to the intervention regimens specified in the study design is often proposed. The motivation for withdrawal of nonadherent participants is that the trial is not a "fair test" of the ideal intervention with these participants included. For example, there may be a few participants in the intervention group who took little or no therapy. If participants do not take their medication, they certainly cannot benefit from it. There could also be participants in the control group who frequently receive the study medication. The intervention and control groups are thus "contaminated." Proponents of withdrawal of nonadherent participants argue that removal of these participants keeps the trial closer to what was intended; that is, a comparison of optimal intervention versus control. The impact of nonadherence on the trial findings is that any observed benefit of the intervention, as compared to the control, will be reduced, making the trial less powerful than it is planned. Newcombe [11], for example, discusses the implication of adherence for the analysis as well as the design and sample size. We discuss this at length in Chap. 8.

A policy of withdrawal from analysis because of participant nonadherence can lead to bias. The overwhelming reason is that participant adherence to a protocol may be related to the intervention. In other words, there may be an interaction between adherence and intervention. Certainly, if nonadherence is greater in one group than another, then withdrawal of nonadherent participants could lead to bias. Even if the frequency of nonadherence is the same for the intervention and control groups, the reasons for nonadherence in each group may differ and may involve different types of participants. The concern would always be whether the same types of participants were withdrawn in the same proportion from each group or whether an imbalance had been created. Of course, an investigator could probably neither confirm nor refute the possibility of bias.

The Coronary Drug Project evaluated several lipid-lowering drugs in people several years after a myocardial infarction. In participants on one of the drugs, clofibrate, total 5-year mortality was 18.2%, as compared with 19.4% in control participants [21, 33]. Among the clofibrate participants, those who had at least 80% adherence to therapy had a mortality of 15%, whereas the poor adherers had a mortality of 24.6% (Table 17.2). This seeming benefit from taking clofibrate was, unfortunately, mirrored in the group taking placebo, 15.1% vs. 28.2%. A similar pattern (Table 17.3) was noted in the Aspirin Myocardial Infarction Study (AMIS) [35]. Overall, no difference in mortality was seen between the aspirin-treated group (10.9%) and the placebo-treated group (9.7%). Good adherers to aspirin had a mortality of 6.1%; poor adherers had a mortality of 21.9%. In the placebo group, the rates were 5.1 and 22%.

Table 17.2 Percent mortality by study group and level of adherence in the Coronary Drug Project

		Drug adherence	
	Overall	≥80%	<80%
Clofibrate	18.2	15.0	24.6
Placebo	19.4	15.1	28.2

Table 17.3 Percent mortality by study group and degree of adherence in the Aspirin Myocardial Infarction Study

	Overall	Good adherence	Poor adherence
Aspirin	10.9	6.1	21.9
Placebo	9.7	5.1	22.0

A trial of antibiotic prophylaxis in cancer patients also demonstrated a relationship between adherence and benefit in both the intervention and placebo groups [40]. Among the participants assigned to intervention, efficacy in reducing fever or infection was 82% in excellent adherers, 64% in good adherers, and 31% in poor adherers. Among the placebo participants, the corresponding figures were 68%, 56%, and 0%.

Another pattern is noted in a three-arm trial comparing two beta-blocking drugs, propranolol and atenolol, with placebo [36]. Approximately equal numbers of participants in each group stopped taking their medication. In the placebo group, adherers and nonadherers had similar mortality: 11.2 and 12.5%, respectively. Nonadherers to the interventions, however, had death rates several times greater than did the adherers: 15.9–3.4% in those on propranolol and 17.6–2.6% in those on atenolol. Thus, even though the numbers of nonadherers in each arm were equal, their risk characteristics as reflected by their morality rates were obviously different.

Pledger [46] provides an analogous example for a schizophrenia trial. Participants were randomized to chlorpromazine or placebo and the 1-year relapse rates were measured. The overall comparison was a 27.8% relapse rate on active medication and 52.8% for those on placebo. The participants were categorized into low or high adherence subgroups. Among the active medication participants, the relapse rate was 61.2% for low adherence and 16.8% for high adherence. However, the relapse rate was 74.7 and 28.0% for the corresponding adherence groups on placebo.

Another example of placebo adherence versus nonadherence is reported by Oakes et al. [45]. A trial of 2,466 heart attack participants compared diltiazem with placebo over a period of 4 years with time to first cardiac event as the primary outcome. Cardiac death or all-cause mortality were additional outcome measures. The trial was initially analyzed according to intention-to-treat with no significant effect of treatment. Qualitative interaction effects were found with the presence or absence of pulmonary congestion which favored diltiazem for patients without pulmonary congestion and placebo in patients with pulmonary congestion. Interestingly, for participants without pulmonary congestion, the hazard ratio or relative risk for time to first cardiac event was 0.92 for those off placebo compared

to those on placebo. For participants with pulmonary congestion, the hazard ratio was 2.86 for participants off placebo compared to those on placebo. For time to cardiac death and to all-cause mortality, hazard ratios exceeded 1.68 in both pulmonary congestion subgroups. This again suggests that placebo adherence is a powerful prognostic indicator and argues for the intention-to-treat analysis.

The definition of nonadherence can also have a major impact on the analysis. This is demonstrated by reanalysis of a trial in breast cancer patients by Redmond et al. [47]. This trial compared a complex chemotherapy regimen with placebo as adjuvant therapy following surgery with disease-free survival as the primary outcome. To illustrate the challenges of trying to adjust analyses for adherence, two measures of adherence were created. Adherence was defined as the fraction of chemotherapy taken while on the study to what was defined by the protocol as a full course. One analysis (Method I) divided participants into good adherers (>85%), moderate adherers (65–84%), and poor adherers (<65%). Using this definition, placebo adherers had a superior disease-free survival than moderate adherers who did better than poor adherers (Fig. 17.1). This pattern of outcome in the placebo group is similar to the CDP clofibrate example. The authors performed a second analysis (Method II) changing the definition of adherence slightly. In this case, adherence was defined as the fraction of chemotherapy taken while on study to what should have been taken while still on study before being taken off treatment for some reason. Note that the previous definition (Method I) compared chemotherapy taken to what would have been taken had the participant survived to the end and adhered perfectly. This subtle difference in definition changed the order of outcome in the placebo group. Here, the poor placebo adherers had the best disease-free survival and the best adherers had a disease-free survival in between the moderate

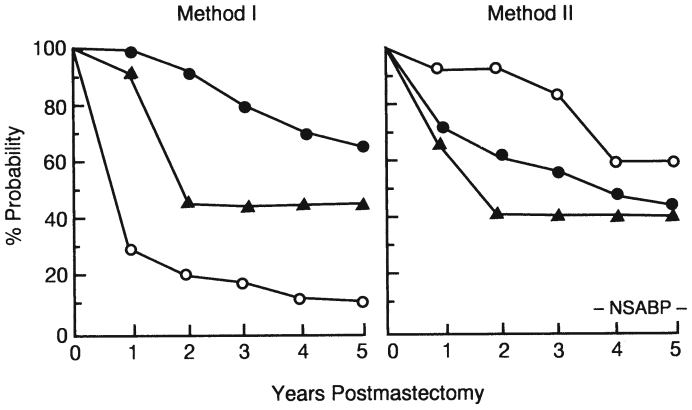


Fig. 17.1 Percentage of disease free survival related to adherence levels of placebo; methods I and II definition of compliance in National Surgical Adjuvant Breast Program (NSABP). Three levels of adherence are: *filled circle* – Good (>85%); *filled triangle* – Moderate (65–84%); *open circle* – Poor (<65%) [47]

and poor adherers. Of special importance is that the participants in this example were all on placebo. Thus, adherence is itself an outcome and trying to adjust one outcome (the primary response variable) for another outcome (adherence) can lead to irrational results.

Detre and Peduzzi have argued that, although as a general rule nonadherent participants should be analyzed according to the study group to which they were assigned, there can be exceptions. They presented an example from the VA coronary bypass surgery trial [37]. In that trial, a number of participants assigned to medical intervention crossed over to surgery. Contrary to expectation, these participants were at similar risk of having an event, after adjusting for a variety of baseline factors, as those who did not crossover. Therefore, the authors argued that the nonadherers should be kept in their original groups but can be censored at the time of crossover. This may be true, but, as seen in the Coronary Drug Project [33], adjustment for known variables does not always account for the observed response. The differences in mortality between adherers and nonadherers remained even after adjustment. Thus, other unknown or unmeasured variables were of critical importance.

Some might think that if rules for withdrawing participants are specified in advance, withdrawals for nonadherence are legitimate. However, the potential for bias cannot be avoided simply because the investigator states, ahead of time, the intention to withdraw participants. This is true even if the investigator is blinded to the group assignment of a participant at the time of withdrawal. Participants were not withdrawn from the analyses in the above examples. However, had a rule allowing withdrawal of participants with poor adherence been specified in advance, the results described above would have been obtained. The type of participants withdrawn would have been different in the intervention and control groups and would have resulted in the analysis of noncomparable groups of adherers. Unfortunately, as noted, the patterns of possible bias can vary and depend on the precise definition of adherence. Neither the magnitude nor direction of that bias is easily assessed or compensated for in analysis.

Adherence is also a response to the intervention. If participant adherence to an intervention is poor compared to that of participants in the control group, widespread use of this therapy in clinical practice may not be reasonably expected. An intervention may be effective, but may be of little value if it cannot be tolerated by a large portion of the participants. For example, in the Coronary Drug Project, the niacin arm showed a favorable trend for mortality, compared with placebo, but niacin caused “hot flashes” and was not easily tolerated [21]. The development of slow release formulations that reduce pharmacologic peaks has lessened the occurrence of side effects.

It is therefore recommended that no participants be withdrawn from analysis in superiority trials for lack of adherence. The price an investigator must pay for this policy is possibly reduced power because some participants who are included in the analysis may not be on optimal intervention. For limited or moderate nonadherence, one can compensate by increasing the sample size, as discussed in Chap. 8, although doing so is costly.

For noninferiority trials, nonadherence will push the two interventions arms to look more alike and thus bias toward the claim of noninferiority. Any attempt to use only adherers in a noninferiority trial, though, will be biased in unknown directions, thus rendering the results uninterpretable. Again, the best policy is to design a trial to have minimum nonadherence, power the trial to overcome non-preventable non-adherence, and then accept the results using the principle of intention-to-treat.

Missing or Poor Quality Data

In many trials, participants may have data missing for a variety of reasons. Perhaps, they were not able to keep their scheduled clinic visits or were unable to perform or undergo the particular procedures or assessments. In some cases, follow-up of the participant was not completed as outlined in the protocol. The challenge is how to deal with missing data or data of such poor quality that they are in essence missing. One approach is to withdraw participants who have poor data completely from the analysis [22, 53, 54]. However, the remaining subset may no longer be representative of the population randomized and there is no guarantee that the validity of the randomization has been maintained in this process.

There is a vast literature on approaches to dealing with missing data [55–65]. Many of these methods assume that the data are missing at random; that is, the reasons the data are missing are not dependent on the measurement that would have been observed. In some contexts, this may be a reasonable assumption. For clinical trials, and clinical research in general, it would be difficult to confirm this assumption. It is, in fact, probably not a valid assumption, as missing data might, for example, be associated with the health status of the participant. Thus, during trial design and conduct, every effort must be made to minimize missing data. If the amount of missing data is relatively small, then the available analytic methods will probably be helpful. If the amount of missing data is substantial, then few, if any, methods will rescue the trial. In this section, we discuss some of the issues that must be kept in mind when analyzing a trial with missing data.

Rubin [58] provided a definition of missing data mechanisms. If data are missing for reasons unrelated to the measurement that would have been observed, then the data are "missing completely at random." If a measure or index allows a researcher to estimate the probability of having missing data, say in a participant with poor adherence to the protocol, then using methods proposed by Rubin and others might allow some adjustment to reduce bias. However, adherence, as indicated earlier, is often associated with a participant's outcome and attempts to adjust for adherence can lead to misleading results.

If participants do not adhere to the intervention and also do not return for follow-up visits, the primary outcome measured may not be obtained unless the outcome is survival or some easily ascertained event. In this situation, an intention-to-treat analysis is not feasible and no analysis is fully satisfactory. Because withdrawal of participants from the analysis is known to be problematic, one approach is to

"impute" or fill in the missing data such that standard analyses can be conducted. This is appealing if the imputation process can be done without adding bias. There are many procedures for imputation. Those based on multiple imputations are probably more robust than single imputation.

A commonly used single imputation method is to carry the last observed value forward. This method, also known as an endpoint analysis, requires the very strong and unverifiable assumption that all future observations, if they were available, would remain constant [46]. Although commonly used, the last observation carried forward method is not generally recommended [64, 65]. Using the average value for all participants with available data, or using a regression model to predict the missing value are alternatives, but in either case, the requirement that the data be missing at random is necessary for proper inference.

A more complex approach is to conduct multiple imputations, typically using regression methods, and then perform a standard analysis for each imputation cycle. The final analysis should take into consideration the variability across the imputation cycles. As with single imputation, the inference based on multiple imputation depends on the assumption that the data are missing at random. Other technical approaches are not described here, but in the context of a clinical trial, none is likely to be satisfactory.

Various other methods for imputing missing values have been described [55–65]. Examples of some of these methods are given by Espeland et al. for a trial measuring carotid artery thickness at multiple anatomical sites using ultrasound [54]. In diagnostic procedures of this type, typically not all measurements can be made. Several imputation methods, based on a mixed effects linear model where regression coefficient and a covariance structure (i.e., variances and correlations), were estimated. Once these were known, this regression equation was the basis for the imputation. Several imputation strategies were used based on different methods of estimating the parameters and whether treatment differences were assumed or not. Most of the imputation strategies gave similar results when the trial data were analyzed. The results indicated up to a 20% increase in efficiency compared to using available data in cross sectional averages.

Imputation techniques such as those described are useful if the data are missing at random; that is, the probability of missing data is not dependent on the measurement that would have been observed or on the preceding measurements. Unfortunately, it is unlikely that data are missing at random. The best that can be offered, therefore, is a series of analyses, each exploring different approaches to the imputation problem. If all, or most, are in general agreement qualitatively, then the results are more persuasive. All analyses should be presented, not just the one with the preferred results.

In long-term trials, participants may be lost to follow-up. In this situation, the status of the participant with regard to any response variable cannot be determined. If mortality is the primary response variable and if the participant fails to return to the clinic, his survival status may still be obtained. If a death has occurred, the date of death can be ascertained. In the Coronary Drug Project [21] where survival experience over 60 months was the primary response variable, four of 5,011 participants

were lost to follow-up (one in a placebo group, three in one treatment group, and none in another treatment group). The Lipid Research Clinics Coronary Primary Prevention Trial [38] followed over 3,800 participants for an average of 7.4 years and was able to assess vital status on all. The Physicians' Health study of over 20,000 US male physicians had complete follow-up for survival status [66]. Many other large simple trials, such as GUSTO [32], have similar nearly complete follow-up experience. Obtaining such low loss to follow-up rates, however, required special effort. In the Women's Health Initiative (WHI), one portion evaluated the possible benefits of hormone replacement therapy (estrogen plus progestin) compared with placebo in postmenopausal women. Of the 16,025 participants, 3.5% were lost to follow-up and did not provide 18 month data [67].

For some conditions, e.g., trials of treatment for substance abuse, many participants fail to return for follow-up visits, and missing data can be 25–30% or even more. Efforts to adjust for missing data must be made, recognizing that biases may very well exist.

An investigator may not be able to obtain any information on some kinds of response variables. For example, if a participant is to have blood pressure measured at the last follow-up visit 12 months after randomization and the participant does not show up for that visit, this blood pressure can never be retrieved. Even if the participant is contacted later, the later measurement does not truly represent the 12-month blood pressure. In some situations, substitutions may be permitted, but in general, this will not be a satisfactory solution. An investigator needs to make every effort to have participants come in for their scheduled visits in order to keep losses to follow-up at a minimum. In the Intermittent Positive Pressure Breathing (IPPB) trial, repeated pulmonary function measurements were required for participants with chronic obstructive pulmonary disease [53]. However, some participants who had deteriorated could not perform the required test. A similar problem existed for the MILIS where infarct size could not be obtained in many of the sickest participants [22].

Individuals with chronic obstructive pulmonary disease typically decline in their pulmonary function and this decline may lead to death, as happened to some participants in the IPPB trial. In this case, the missing data were not missing at random and censoring was said to be informative. One simple method for cases such as the IPPB study is to define a decreased performance level considered to be a clinical event. Then the analysis can be based on time to the clinical event of deterioration or death, incorporating both pieces of information. Survival analysis, though, assumes that loss of follow-up is random and independent of risk of the event. Methods relaxing the missing at random assumption have been proposed [68, 69], but require other strong assumptions, the details of which are beyond the scope of this text.

If the number of participants lost to follow-up differs in the study groups, the analysis of the data could be biased. For example, participants who are taking a new drug that has side effects may, as a consequence, miss scheduled clinic visits. Events may occur but be unobserved. These losses to follow-up would probably not be the same in the control group. In this situation, there may be a bias favoring the new drug. Even if the number lost to follow-up is the same in each study group, the possibility

of bias still exists because different reasons may be involved. The participants who are lost in each group may have quite different prognoses and outcomes.

An example of differential follow-up was reported by the Comparison of Medical Therapy, Pacing, and Defibrillation in Chronic Heart Failure (COMPANION) trial [70]. COMPANION compared a cardiac pacemaker or a pacemaker plus defibrillator with best pharmacologic treatment in people with chronic heart failure. Over 1,500 participants were randomized. Two primary outcomes were assessed; death and death plus hospitalization. Individuals randomized to one of the device arms did not know to which device they had been assigned, but those on the pharmacologic treatment arm were aware that no device had been installed. During the course of the trial, the pacemaker plus defibrillator devices, made by two different manufacturers, were approved by a regulatory agency. As a result, participants in the pharmacologic treatment arm began to drop-out from the trial and some also withdrew their consent. Many requested one of the newly approved devices. Thus, when the trial was nearing completion, the withdrawal rate was 26% in the pharmacologic treatment arm and 6–7% in the device arms. Additionally, no further follow-up information could be collected on those who withdrew consent. Clearly, censoring at the time of withdrawal was not random and the possibility that it was related to disease status could not be ruled out, thus creating the possibility of serious bias. This situation could have jeopardized an otherwise well designed and conducted trial in people with a serious medical condition. However, the investigators initiated a complicated process of reconsenting the participants to allow for collection of the primary outcomes. After completing this process, assessment of the status for death plus hospitalization and vital status were 91 and 96%, respectively, in the pharmacologic treatment group. Outcome ascertainment for the two device arms was 99% or better. The final results demonstrated that both the pacemaker and the defibrillator plus pacemaker reduced death plus hospitalization and further that the defibrillator plus pacemaker reduced mortality. These results were important in the treatment of chronic heart failure. However, not correcting for the initial differential loss to follow-up would have rendered the COMPANION trial data perhaps uninterpretable. In Fig. 17.2, the Kaplan–Meier curves for mortality for the two intervention arms are provided with the most complete data available.

Often, protocol designs call for follow-up to terminate at some period, for example 7, 14, or 30 days, after a participant has stopped adhering to his or her intervention, even though the intended duration of intervention would not have ended. The concept is that “off intervention” means “off study”; i.e., assessment for nonadherent participants ends when intervention ends. We do not endorse this concept. Although time to event analysis may be censored at the time of last follow-up, going off intervention or control is not likely random and may be related to participant health status. Important events, including serious adverse events, may occur beyond the follow-up period and might be related to the intervention. As noted above, though, survival analysis assumes that censoring is independent of the primary event. The practice of not counting events at the time of, or shortly after, intervention discontinuation is all too common, and can lead to problems in the

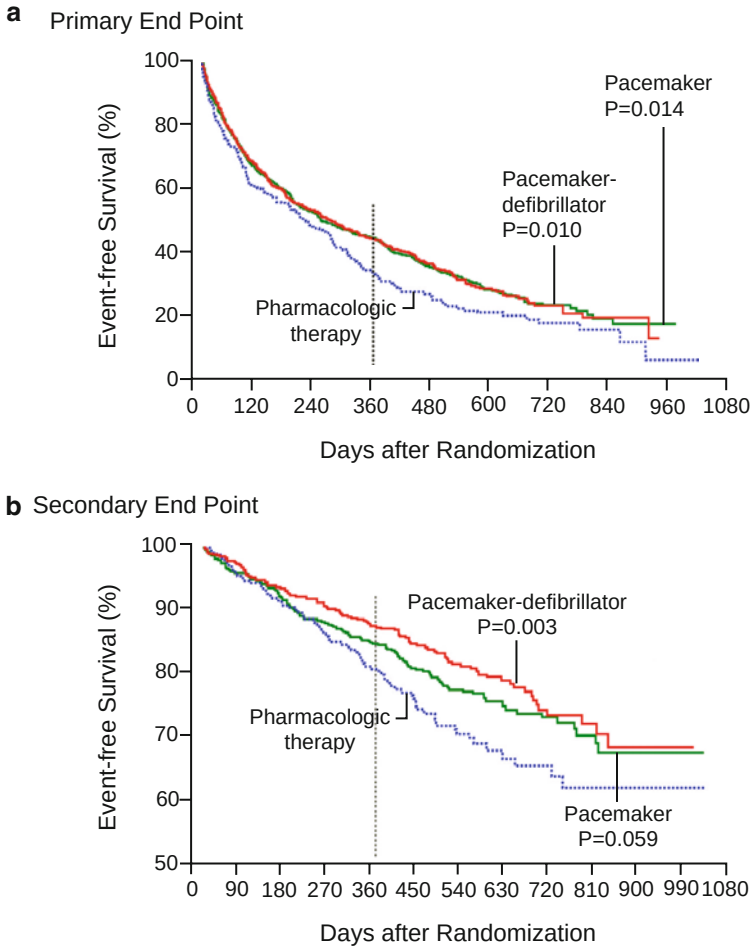


Fig. 17.2 In the COMPANION trial, Kaplan–Meier estimates of (**a**) the time to primary end point of death or hospitalization for any cause (**b**) the time to the secondary end point of death from any cause [70]. Reprinted with the permission of the Massachusetts Medical Society, copyright © 2004, all rights reserved.

interpretation of the final results. An instructive example is the Adenomatous Polyp Prevention on Vioxx (APPROVe) trial [71]. This randomized double-blind trial compared a cyclo-oxygenase (COX)-II inhibitors with placebo in people with a history of colorectal adenomas. Previous trials of COX-II inhibitors had raised concern regarding long-term cardiovascular risk. Thus, while the APPROVe trial was a cancer prevention trial, attention also focused on cardiovascular events, in particular thrombotic events and cardiovascular death, nonfatal myocardial infarction, and nonfatal stroke. However, participants who stopped taking their medication during the trial were not followed beyond 14 days after the time of

discontinuation. The Kaplan–Meier cardiovascular risk curve is shown in Fig. 17.3. Note that for the first 18 months, the two curves are similar and then begin to diverge. Controversy arose as to whether there was an 18-month lag time in the occurrence of cardiovascular events for this particular COX-II inhibitor [72, 73].

Due to the controversy, the investigators and sponsor launched an effort to collect information on all trial participants for up to a year beyond the close of the trial. This extended follow-up, referred to here as APPROVe+1, was able to collect selected cardiovascular events of nonfatal myocardial infarction, nonfatal stroke, and cardiovascular death [74], as shown in Fig. 17.4. The time to event curves begin to separate from the beginning and continue throughout the extended follow-up, with a hazard ratio of 1.8 ($p = 0.006$). There was a corresponding nonsignificant increase in mortality.

Censoring follow-up when participants go off their intervention is a common error that leads to problems like those encountered by the APPROVe trial. Going off intervention, and thus censoring follow-up at some number of days afterwards, is not likely to be independent of the disease process or how a participant is doing. At least, it would be difficult to demonstrate such independence. Yet, survival analysis

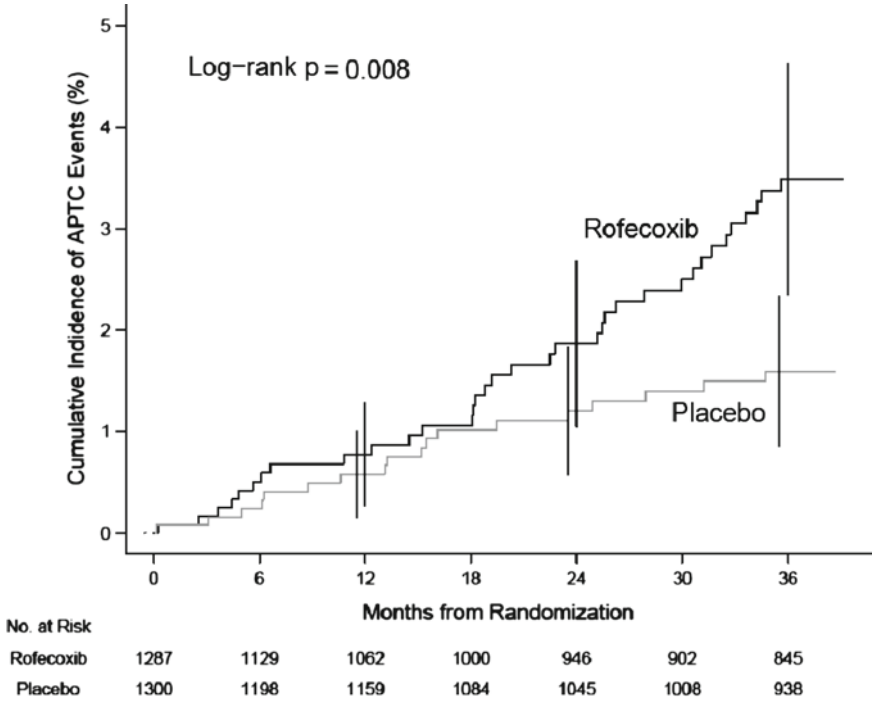


Fig. 17.3 APPROVe Kaplan–Meier estimates of time to event from the AntiPlatelet Trialists' Collaborative (APTC) outcomes (death from cardiovascular causes, nonfatal myocardial infarction or nonfatal stroke) with censoring 14 days after participants stopped therapy [74]. Reproduced with the permission of Elsevier Ltd. for *Lancet*

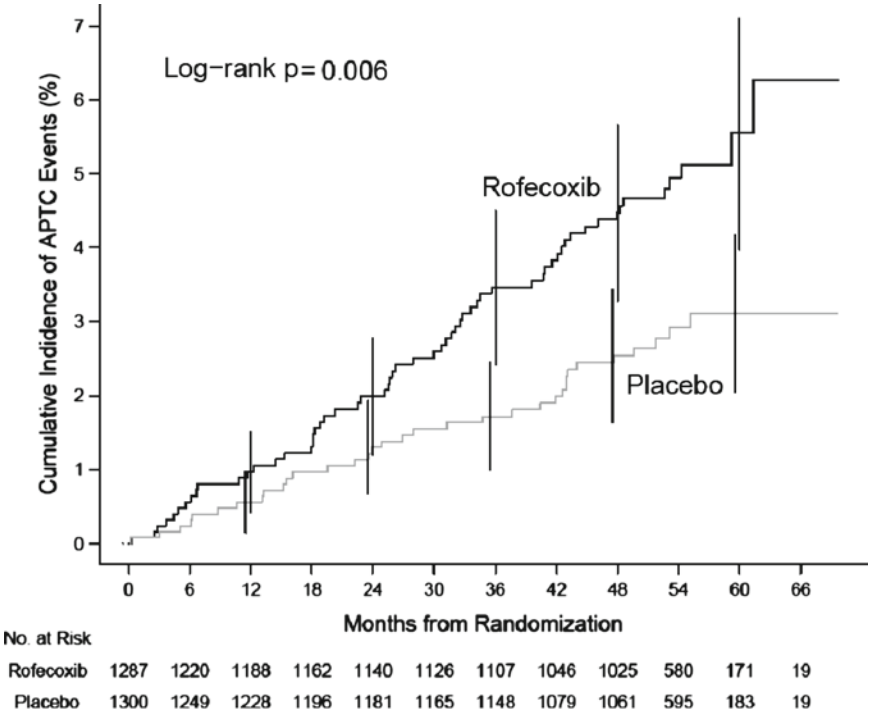


Fig. 17.4 APPROVe Kaplan–Meier estimates of time to event for the AntiPlatelet Trialists' Collaborative (APTC) outcome (death from cardiovascular causes, nonfatal myocardial infarction or nonfatal stroke) counting all events observed for an additional year of follow-up after the trial was initially terminated [74]. Reproduced with the permission of Elsevier Ltd. for *Lancet*

and most other analyses assume that the censoring is independent. The principle lesson here is that "off intervention should not mean off study."

An outlier is an extreme value significantly different from the remaining values. The concern is whether extreme values in the sample should be excluded from the analysis. This question may apply to a laboratory result, to the data from one of several areas in a hospital or from a clinic in a multicenter trial. Removing outliers is not recommended unless the data can be clearly shown to be erroneous. Even though a value may be an outlier, it could be correct, indicating that on occasions an extreme result is possible. This fact could be very important and should not be ignored. Long ago, Kruskal [75] suggested carrying out an analysis with and without the "wild observation." If the conclusions vary depending on whether the outlier values are included or excluded, one should view any conclusion cautiously. Procedures for detecting extreme observations have been discussed [76–79], and the publications cited can be consulted for further detail.

An interesting example given by Canner et al. [78] concerns the Coronary Drug Project. The authors plotted the distributions of four response variables for each of the 53 clinics in that multicenter trial. Using total mortality as the

response variable, no clinics were outlying. When nonfatal myocardial infarction was the outcome, only one clinic was an outlier. With congestive heart failure and angina pectoris, response variables which are probably less well defined, there were nine and eight outlying clinics, respectively.

Competing Events

Competing events are those that preclude the assessment of the primary response variable. They can reduce the power of the trial by decreasing the number of participants available for follow-up. If the intervention can affect the competing event, there is also the risk of bias. In some clinical trials, the primary response variable may be cause-specific mortality, such as death due to myocardial infarction or sudden death, rather than total mortality [80–83]. The reason for using cause-specific death as a response variable is that a therapy often has specific mechanisms of action that may be effective against a disease or condition. In this situation, measuring death from all causes, most of which are not likely to be affected by the intervention, can "dilute" the results. For example, a study drug may be antiarrhythmic and thus sudden cardiac death might be the selected response variable. Other causes of death such as cancer and accidents would be competing events.

Even if the response variable is not cause-specific mortality, death may be a factor in the analysis. This is particularly an issue in long term trials in the elderly or high risk populations. If a participant dies, future measurements will be missing. Analysis of nonfatal events in surviving participants has the potential for bias, especially if the mortality rates are different in the two groups.

In a study in which cause-specific mortality is the primary response variable, deaths from other causes are treated statistically as though the participants were lost to follow-up from the time of death (Chap. 15), and these deaths are not counted in the analysis. In this situation, the analysis, however, must go beyond merely examining the primary response variable. An intervention may or may not be effective in treating the condition of interest but could be harmful in other respects. Therefore, total mortality should be considered as well as cause-specific fatal events. Similar considerations need to be made when death occurs in studies using nonfatal primary response variables. This can be done by considering tables that show the number of times the individual events occur, one such event per person. No completely satisfactory solution exists for handling competing events. At the very least, the investigator should report all major outcome categories; for example, total mortality, as well as cause-specific mortality and morbid events.

In many cases, there may be recurring events. Many trials simply evaluate the time to the first event and do not count the additional events in the time to event analysis. Tables may show the total number of events in each intervention arm. Some attempts to further analyze recurrent events have been made, using for example the data from the COMPANION trial [70, 82]. These methods are complicated, however, and will not be covered in this text.

Composite Outcomes

In recent years, many trials have used combinations of clinical and other outcomes as a composite response variable [80–83]. One major motivation is to increase the event rate and thus reduce the sample size that might otherwise have been required had just one of the components been selected as the primary outcome. Another motivation is to combine events that have a presumed common etiology and thus get an overall estimate of effect. The sample size is usually not based on any single component.

There are challenges in using a composite outcome [84, 85]. The components may not have equal weight or clinical importance, especially as softer outcomes are added. The components may go in opposite directions or at least not be consistent in indicating intervention effect. One component may dominate the composite. Results with any single component are based on a smaller number of events and thus the power for that component is greatly reduced. Rarely do we find significance for a component, nor should we expect it in general. Regardless of the composition of the composite, analyses should be conducted for each component, or in some cascading sequence. For example, if the composite were death, myocardial infarction, stroke or heart failure hospitalization, the analysis sequence might be death, death plus myocardial infarction or stroke, and death plus hospitalization. The reason for including death is to take into account competing risk of death for the other components, in addition to its obvious clinical importance.

As pointed out in Chap. 3, it is essential that follow-up continue after the first event in the composite outcome occurs. Analysis will include looking at the contribution of each component to the overall but should also include time to event for each component separately. As indicated, if follow-up does not continue, only partial results are available for each component and analysis of those events separately could be misleading.

There are several examples where the use of a composite such as death, myocardial infarction, and stroke has been used as a primary or leading secondary outcome [28–31]. These outcomes are all clinically relevant. In these trials, the outcomes all trended in the same direction. However, that is not always the case.

In the Pravastatin or Atorvastatin Evaluation and Infection Therapy (PROVE IT) trial, the 80 mg atorvastatin arm was more effective than the 40 mg pravastatin arm in reducing the incidence of death, myocardial infarction, stroke, required hospitalization due to unstable angina and revascularization [83]. Stroke results, one of the key components, went in the opposite direction. These results complicate the interpretation. Should investigators think that the atorvastatin improves the composite or just those components that are in the same direction as the composite? As would be expected, the differences for the components were not, in themselves, statistically significant.

Another interesting example is provided by the WHI, which was a large factorial design trial in postmenopausal women [67]. As discussed earlier and in Chap. 16, one part involved hormone replacement therapy which contained two strata, women with a uterus and those without. Women with a uterus received either estrogen plus progestin or matching placebo; those without a uterus received estrogen alone or a matching placebo. Due to the multiple actions of hormone replacement therapy, one response variable was a global outcome mortality, coronary heart disease, bone loss reflected by hip fracture rates, breast cancer, colorectal cancer, pulmonary embolism, and stroke. As seen in Fig. 16.6b, for the estrogen plus progestin stratum, there was essentially no effect on mortality and a small but nonsignificant effect in the global index, when compared to placebo. However, as shown in Fig. 16.6a, the various components went in different directions. Hip fracture and colorectal cancer had a favorable response to hormone replacement therapy. Pulmonary embolism, stroke, and coronary heart disease went in an unfavorable direction. Thus, any interpretation of the global index, which is a composite, requires careful examination of the components. Of course, few trials would have been designed with adequate power for the individual components so that the interpretation must be qualitative, looking for consistency and biological plausibility.

Experience suggests that composite outcome variables should be used cautiously and only include those components that have relatively equal clinical importance, frequency, and anticipated response to the presumed mechanism of action of the intervention [84]. As softer and less relevant outcomes are added, the interpretation becomes less clear, particularly if the less important component occurs more frequently than others, driving the overall result. Significance by any individual component cannot be expected, but there should be a plausible consistency across the components.

Covariate Adjustment

The goal in a clinical trial is to have study groups that are comparable except for the intervention being studied. Even if randomization is used, all of the prognostic factors may not be perfectly balanced, especially in smaller studies. Even if no prognostic factors are significantly imbalanced in the statistical sense, an investigator may, nevertheless, observe that one or more factors favor one of the groups. In either case, covariate adjustment can be used in the analysis to minimize the effect of the differences. However, covariate adjustment is not likely to eliminate the effect of these differences. Covariance analysis for clinical trials has been reviewed in numerous articles [86–107].

Adjustment also reduces the variance in the test statistic. If the covariates are highly correlated with outcome, this can produce more sensitive analyses. The specific adjustment procedure depends on the type of covariate being adjusted for and the type of response variable being analyzed. If a covariate is discrete, or if a continuous variable is converted into intervals and made discrete, the analysis is sometimes

referred to as "stratified." A *stratified analysis*, in general terms, means that the study participants are subdivided into smaller, more homogeneous groups, or strata. A comparison of study groups is made within each stratum and then averaged over all strata to achieve a summary result for the response variable. This result is adjusted for group imbalances in the discrete covariates. If a response variable is discrete, such as the occurrence of an event, the stratified analysis might take the form of a Mantel–Haenszel statistic described briefly in the Appendix to this chapter.

If the response variable is continuous, the stratified analysis is referred to as *analysis of covariance*. This uses a model which, typically, is linear in the covariates. A simple example for a response Y and covariate X would be $Y = \alpha_j + \beta(X - \mu) + error$ where β is a coefficient representing the importance of the covariate X and is assumed to be the same in each group, μ is the mean value of X, and α_j is a parameter for the contribution of the overall response variable jth group (e.g., $j = 1$ or 2). The basic idea is to adjust the response variable Y for any differences in the covariate X between the two groups. Under appropriate assumptions, the advantage of this method is that the continuous covariate X does not have to be divided into categories. Further details can be found in statistics textbooks [1–8]. If time to an event is the primary response variable, then survival analysis methods that allow for adjustments of discrete or continuous covariates may be used [100]. However, whenever models are employed, the investigator must be careful to evaluate the assumptions required and how closely they are met. Analysis of covariance can be attractive, but may be abused if linearity is assumed when the data are nonlinear, if the response curve is not parallel in each group, or if assumptions of normality are not met [90]. If measurement errors in covariates are substantial, the lack of precision can be increased [99]. For all of these reasons, covariate adjustment models may be useful in the interpretation of data, but should not be viewed as absolutely correct.

Regardless of the adjustment procedure, covariates should be measured at baseline. Except for certain factors such as age, sex, or race, any variables that are evaluated after initiation of intervention should be considered as response variables. Group comparisons of the primary response variable, adjusted for other response variables, are discouraged. Interpretation of such analyses is difficult because group comparability may be lost.

Surrogates as Covariates

Adjustment for various surrogate outcomes may be proposed. In a trial of clofibrate [105], the authors reported that those participants who had the largest reduction in serum cholesterol had the greatest clinical improvement. However, reduction in cholesterol is probably highly correlated with adherence to the intervention regimen. Since, as discussed earlier, adherers in one group may be different from adherers in another group, analyses that adjust for a surrogate for adherence can be biased.

This issue was addressed in the Coronary Drug Project [33]. Adjusted for baseline factors, the 5-year mortality was 18.8% in the clofibrate group ($N=997$) and 20.2% in the placebo group ($N=2{,}535$), an insignificant difference. For participants with baseline serum cholesterol greater than or equal to 250 mg/dl, the mortality was 17.5 and 20.6% in the clofibrate and placebo groups, respectively. No difference in mortality between the groups was noted for participants with baseline cholesterol of less than 250 mg/dl (20.0% vs. 19.9%). Those participants with lower baseline cholesterol in the clofibrate group who had a reduction in cholesterol during the trial had 16.0% mortality, as opposed to 25.5% mortality for those with a rise in cholesterol (Table 17.4). This would fit the hypothesis that lowering cholesterol is beneficial. However, in those participants with high baseline cholesterol, the situation was reversed. An 18.1% mortality was seen in those who had a fall in cholesterol, and a 15.5% mortality was noted in those who had a rise in cholesterol. The best outcome, therefore, appeared to be in participants on clofibrate whose low baseline cholesterol dropped or whose high baseline cholesterol increased. As seen earlier, adherence can affect outcomes in unexpected ways, and the same is true of surrogates for adherence.

Modeling the impact of adherence on a risk factor and thus on the response has also received attention [89, 93]. Regression models have been proposed that attempt to adjust outcome for the amount of risk factor change that could have been attained with optimum adherence. One example of this was suggested by Efron and Feldman [89] for a lipid research study. However, Albert and DeMets [93] showed that these models are very sensitive to assumptions about the independence of adherence and health status or response. If these assumptions using these regression models are violated, uninterpretable results emerge, such as that for the chlofibrate and serum cholesterol example described above.

Clinical trials of cancer treatment commonly analyze results by comparing responders to nonresponders [86, 87]. That is, those who go into remission or have a reduction in tumor size are compared to those who do not. One early survey indicated that such analyses were done in at least 20% of published reports [90]. The authors of that survey argued that statistical problems, due to lack of random assignment, and methodological problems, due both to classification of response and inherent differences between responders and nonresponders, can occur. These will often yield misleading results, as shown by Anderson et al. [87]. They pointed out that participants "who eventually become responders must survive long enough to be evaluated as responders." This factor can invalidate some statistical tests that compare responders

Table 17.4 Percent 5-year mortality in the clofibrate group, by baseline cholesterol and change in cholesterol in the Coronary Drug Project

	Baseline cholesterol	
	<250 mg/dl	≥250 mg/dl
Total	20.0	17.5
Fall in cholesterol	16.0	18.1
Rise in cholesterol	25.5	15.5

to nonresponders. Those authors present two statistical tests that avoid bias. They note, though, that even if the tests indicate a significant difference in survival between responders and nonresponders, it cannot be concluded that increased survival is due to tumor response. Thus, aggressive intervention, which may be associated with better response, cannot be assumed to be better than less intensive intervention, which may be associated with poorer response. Anderson and colleagues state that only a truly randomized comparison can say which intervention method is preferable. What is unsaid, and illustrated by the Coronary Drug Project examples, is that even comparison of good responders in the intervention group with good responders in the control can be misleading, because the reasons for good response may be different.

Morgan [44] provided a related example of comparing duration of response in cancer patients, where duration of response is the time from a favorable response such as tumor regression (partial or total) to remission. This is another form of defining a subgroup of posttreatment outcome, that is, tumor response. In a trial comparing two complex chemotherapy regiments (*A* vs. *B*) in small cell lung cancer, the tumor response rates were 64 and 85%, with median duration of 245 days and 262 days, respectively. When only responders were analyzed, the slight imbalance in prognostic factors was substantially increased. Extensive disease was evident at baseline in 48 and 21% of the two treatment responder groups. Thus, while it may be theoretically possible to adjust for prognostic factors, in practice, such adjustment may decrease bias, but will not eliminate it. Because not all prognostic factors are known, any model is only an approximation to the true relationship.

The Cox proportional hazards regression model for the analysis of survival data (Chap. 15) allows for covariates in the regression to vary with time [88]. This has been suggested as a way to adjust for factors such as adherence and level of response. It should be pointed out that, like simple regression models, this is vulnerable to the same biases described earlier in this chapter. For example, if cholesterol level and cholesterol reduction in the CDP example were used as time dependent covariates in the Cox model, the estimator of treatment effect would be biased due to the effects shown in Table 17.4.

Rosenbaum [92] provides a nice overview of adjustment for concomitant variables that have been affected by treatment in both observational and randomized studies. He states that "adjustments for posttreatment concomitant variables should be avoided when estimating treatment effects except in rather special circumstances, since adjustments themselves can introduce a bias where none existed previously."

A number of additional methodologic attempts to adjust for adherence have also been conducted. Newcombe [11], for example, suggested adjusting estimates of intervention effect on the extent of nonadherence. Robins and Tsiatis [106] proposed a causal inference model. Lagakos et al. [43] evaluated censoring survival time, or time to an event, at the point when treatment is terminated. The rationale is that participants are no longer able to completely benefit from the therapy. However, the hazard ratio estimated by this approach is not the hazard that would have been estimated if participants had not terminated treatment. The authors stated that it is not appropriate to evaluate treatment benefit by comparing the hazard rates estimated by censoring for treatment termination [43].

Baseline Variables as Covariates

The issue of stratification was first raised in the discussion of randomization (Chap. 6). For large studies, the recommendation was that stratified randomization is usually unnecessary because overall balance would nearly always be achieved and that stratification would be possible in the analysis. For smaller studies, baseline adaptive methods could be considered but the analysis should include the covariates used in the randomization. In a strict sense, analysis should always be stratified if stratification was used in the randomization. In such cases, the adjusted analysis should include not only those covariates found to be different between the groups, but also those stratified during randomization. Of course, if no stratification is done at randomization, the final analysis is less complicated since it would involve only those covariates that turn out to be imbalanced or to be of special interest associated with the outcomes.

As stated in Chap. 6, randomization tends to produce comparable groups for both measured and unmeasured baseline covariates. However, not all baseline covariates will be closely matched. Adjusting treatment effect for these baseline disparities continues to be debated. Canner [96] describes two points of view, one which argues that "if done at all, analyses should probably be limited to covariates for which there is a disparity between the treatment groups and that the unadjusted measure is to be preferred." The other view is "to adjust on only a few factors that were known from previous experience to be predictive." Canner [96], as well as Beach and Meier [94], indicate that even for moderate disparity in baseline comparability, or even if the covariates are moderately predictive, it is possible for covariate adjustment to have a nontrivial impact on the measure of treatment effect. However, Canner [96] also points out that it is "often possible to select specific covariates out of a large set in order to achieve a desired result." In addition, he shows that this issue is true for both small and large studies. For this reason, it is critical that the process for selecting covariates be specified in the protocol and adhered to in the primary analyses. Other adjustments may be used in exploratory analyses.

Another issue is testing for *covariate interaction* in a clinical trial [95, 97, 101, 102, 107]. Treatment-covariate interaction is defined when the response to treatment varies according to the value of the covariate [95]. Peto [107] defines treatment covariate interactions as quantitative or qualitative. Quantitative interactions indicate that the magnitude of treatment effect varies with the covariate but still favors the same intervention (Fig. 17.5a). Qualitative interaction involves a change in the better intervention as the covariate changes in value (Fig. 17.5b). A quantitative interaction, for example, would be if the benefit of treatment for blood pressure on mortality varied in degree by the level of baseline blood pressure but still favoring the same intervention (See Fig. 17.5a). A qualitative interaction would exist if lowering blood pressure was beneficial for severe hypertension, but less beneficial or even harmful for mild hypertension. Intervention effects can vary by chance across

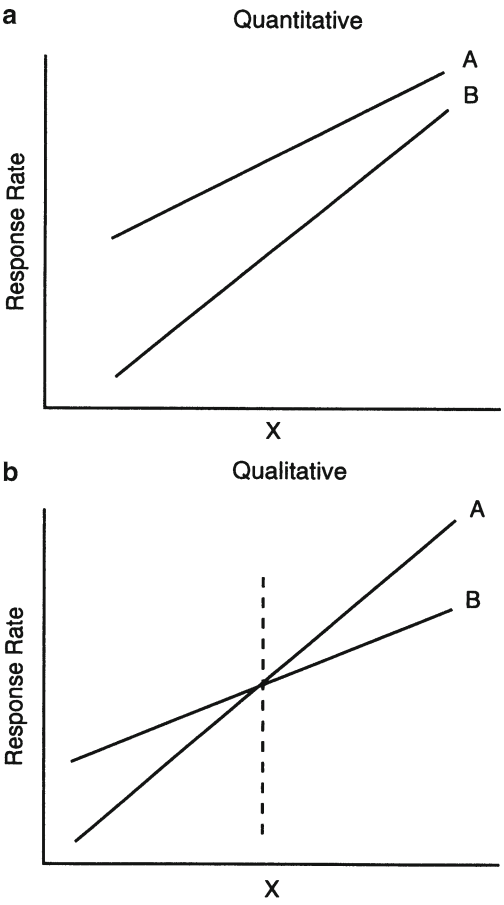


Fig. 17.5 Two types of intervention–covariate interactions [107]

levels of the covariate, even changing direction, so a great deal of caution must be taken in the interpretation. One can test formally for interaction, but requiring a significant interaction test is much more cautious than reviewing the magnitude of intervention effect within each subgroup. Byar [95] presents a nice illustration example shown in Table 17.5. Two treatments, A and B, are being compared by the difference in mean response, $Y = \overline{X}_A - \overline{X}_B$, and S is the standard error of Y. In the upper panel, the interaction test is not significant, but examination of the subgroups is highly suggestive of interaction. The lower panel is more convincing for interaction, but we still need to examine each subgroup to understand what is going on.

Methods have been proposed for testing for overall interactions [101, 102]. However, Byar's concluding remarks [95] are noteworthy when he says,

> one should look for treatment-covariate interactions, but, because of the play of chance in multiple comparisons, one should look very cautiously in the spirit of exploratory data analysis rather than that of formal hypothesis testing. Although the newer statistical methods may help decide whether the data suffice to support a claim of qualitative interactions and permit a more precise determination of reasonable p values, it seems to me unlikely that these methods will ever be as reliable a guide to sensible interpretation of data as will medical plausibility and replication of the findings in other studies. We are often warned to specify the interactions we want to test in advance in order to minimize the multiple comparisons problem, but this is often impossible in practice and in any case would be of no help in evaluating unexpected new findings. The best advice remains to look for treatment-covariate interactions but to report them skeptically as hypotheses to be investigated in other studies.

As indicated in Chap. 6, the randomization in multicenter trials should be stratified by clinic. The analysis of such a study should, strictly speaking, incorporate the clinic as a stratification variable. Furthermore, the randomization should be blocked in order to achieve balance over time in the number of participants randomized to each group. These "blocks" are also strata and, ideally, should be included in the analysis as a covariate. However, there could be a large number of strata, since there may be many clinics and the blocking factor within any clinic is usually anywhere from four to eight participants. The use of these blocking covariates is probably not necessary in the analysis. Some efficiency will be lost for the sake of simplicity, but the sacrifice should be small.

As Fleiss [12] describes, clinics differ in their demography of participants and medical practice as well as adherence to all aspects of the protocol. These factors are likely to lead to variation in treatment response from clinic to clinic. In the BHAT [20], most, but not all, of the 30 clinics showed a trend for mortality benefit from propranolol. A few indicated a negative trend. In the AMIS [103], data from a few clinics suggested a mortality benefit from aspirin although most clinics indicated little or no benefit. Most reported analyses probably do not stratify by clinic, but simply combine the results of all clinics. However, at least one of the primary analyses should average within-clinic differences, an analysis that is always valid, even in the presence of clinic-treatment interaction [101].

Table 17.5 Examples of apparent treatment-covariate interactions [95]. Let $Y = \overline{X}_A - \overline{X}_B$

	Statistic	*SE* of *Y*	*P* value (2 tail)
Unconvincing			
Overall test	$Y=2S$	S	0.045
Subsets	$Y_1=3S$	$S\sqrt{2}$	0.034
	$Y_2=1S$	$S\sqrt{2}$	0.480
Interaction	$Y_1-Y_2=2S$	$2S$	0.317
More convincing			
Overall test	$Y=2S$	S	0.045
Subsets	$Y_1=4S$	$S\sqrt{2}$	0.005
	$Y_2=0$	$S\sqrt{2}$	1.000
Interaction	$Y_1-Y_2=4S$	$2S$	0.045

Subgroup Analyses

While covariance or stratified analysis adjusts the overall comparison of main outcomes for baseline variables, another common analytic technique is to subdivide or subgroup the enrolled participants [108–123]. Here, the investigator looks specifically at the intervention-control comparison within one or more particular subgroups rather than the overall comparison. One of the most frequently asked questions during the design of a trial and when the results are analyzed is, "Among which group of participants is the intervention most beneficial or harmful?" It is important that subgroups be examined. Clinical trials require considerable time and effort to conduct, and the resulting data deserve maximum evaluation. The hope is to refine the primary hypothesis and specify to whom, if anyone, the intervention should be recommended. Nevertheless, care must be exercised in the interpretation of subgroup findings. As the number of subgroups increases, the potential for chance findings increases due to multiple comparisons [123]. As discussed earlier in this chapter, categorization of participants by any outcome variable, e.g., adherence, can lead to biased conclusions. Only baseline factors are appropriate for use in defining subgroups.

Subgroups may be identified in several ways that affect the strength of their results. First, subgroup hypotheses may be specified in the study protocol. Because these are defined in advance, they have the greatest credibility. There is likely to be, however, low power for detecting differences in subgroups. Therefore, investigators should not pay as much attention to statistical significance for subgroup questions as they do for the primary question. Recognizing the low chance of seeing significant differences, descriptions of subgroup effects are often qualitative. On the other hand, testing multiple questions can increase the chance of a Type I error. Therefore, if one were to perform tests of significance on a large number of subgroup analyses, there will be an increased probability of false positive results unless adjustments are made. Even in this scenario, there are reasons to be cautious as illustrated by the Prospective Randomized Amlodipine Survival Evaluation Study (PRAISE), a large multicenter trial [115]. In this trial, participants with chronic heart failure were stratified by ischemic and nonischemic etiology. While the primary outcome of death or cardiovascular hospitalization was nonsignificant and the secondary outcome of overall survival outcome was nearly significant ($p = 0.07$), almost all of the risk reduction was in the nonischemic subgroup. The risk reduction was 31% for the primary outcome ($p = 0.04$) and 46% for mortality ($p < 0.001$). However, the more favorable result was expected to be in the ischemic subgroup, not the nonischemic subgroup. Thus, the investigators recommended that a second trial be conducted in the patient population with nonischemic chronic heart failure using a nearly identical protocol to confirm this impressive subgroup result [115]. The results of the PRAISE-II trial proved disappointing with no reduction in either the primary or secondary outcome [116]. Thus, the previous predefined subgroup result could not be confirmed.

Some subgroups may be implied, but may not be explicitly stated in the protocol. For example, if randomization is stratified by age, sex, or stage of disease, it might

be reasonable to infer that subgroup hypotheses related to those factors were in fact considered in advance. Of course, the same problems in interpretation apply here as with formally prespecified subgroups.

A third type of analysis concerns subgroups identified by other, similar trials. If one study reports that the observed difference between intervention and control appears to be concentrated in a particular subgroup of participants, it is appropriate to see if the same findings occur in another trial, even though that subgroup was not prespecified. Problems here include comparability of definition. It is unusual for different trials to have baseline information sufficiently similar to allow for characterization of identical subgroups.

On occasion, during the monitoring of a trial, particular subgroup findings may emerge and be of special interest. If additional participants remain to be enrolled into the trial, one approach is to test the new subgroup hypothesis in the later participants. With small numbers of participants, it is unlikely that significant differences will be noted. If, however, the same pattern emerges in the newly created subgroup, the hypothesis is considerably strengthened.

The weakest type of subgroup analysis involves posthoc analysis, sometimes referred to as "data-dredging" or "fishing." Such analysis is determined by the data themselves. Because many comparisons are theoretically possible, tests of significance become difficult to interpret and should be challenged. Such analyses should serve primarily to generate hypotheses for evaluation in other trials. An example of subgrouping that was challenged comes from a study of diabetes in Iceland. Male children under the age of 14 and born in October were claimed to be at highest risk of ketosis-prone diabetes. Goudie [121] challenged whether the month of October emerged from poststudy analyses biased by knowledge of the results. The ISIS-2 trial [114] illustrated a spurious subgroup finding that suggested treatment benefit of aspirin after myocardial infarction was not present in individuals born under Gemini or Libra astrological signs. A similar example [118] suggests twice as many participants with bronchial carcinoma were born in the month of March ($p < 0.01$) although this observation could not be reproduced [119, 120].

A number of trials of beta-blocking drugs were conducted in people who had a myocardial infarction. One found that the observed benefit was restricted to participants with anterior infarctions [108]. Another claimed improvement only in participants 65 years or younger [109]. In the BHAT, it was observed that the greatest relative benefit of the intervention was in participants with complications during the infarction [110]. These subgroup findings however, were not consistently confirmed in other trials [113].

An interesting posthoc subgroup analyses was reported by the Metoprolol CR/XL Randomized Intervention Trial (MERIT) [122]. This trial, which evaluated the effect of a beta-blocker in participants with chronic heart failure, had two primary outcomes. One was all-cause mortality and the other was death plus hospitalization. MERIT was terminated early by the monitoring committee due to a highly significant reduction in mortality, as shown in Fig. 17.6a, and similar reductions in death plus

hospitalization. The results are remarkably consistent across all of the predefined subgroups for mortality, mortality plus hospitalization, and mortality plus heart failure hospitalization as shown in Fig. 17.6b. Moreover, the results were very consistent with those from two other beta-blocker trials [28, 30], as shown in Fig. 17.7a, b. However, post hoc analyses during review by regulatory agencies compared results among countries. These results are shown in Fig. 17.6c. Of note is that for mortality, the relative risk in the United States slightly favors placebo, in contrast to the mortality results for the trial as a whole. With respect to outcomes of mortality plus hospitalization, and mortality and hospitalization for heart failure, the U.S. data are consistent with the overall trial results. As noted by Wedel et al. [122], the analysis for interaction depends on how the regional subgroups are formed. Whether the observed regional difference is due to chance or real has been debated, but Wedel and colleagues argued that is not consistent with other external data, not internally consistent within MERIT and not biologically plausible, and thus is most likely due to chance. This result does, however, point out the risks of posthoc subgroup analyses.

Regardless of how subgroups are selected, several factors can provide supporting evidence for the validity of the findings. As mentioned, similar results obtained in several studies strengthen interpretation. Internal consistency within a study is also a factor. If similar subgroup results are observed at most of the sites of a multicenter trial, they are more likely to be true. Plausible, post hoc biological

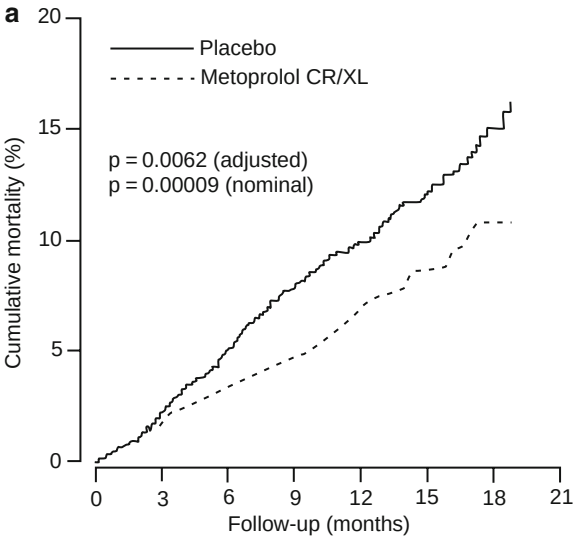
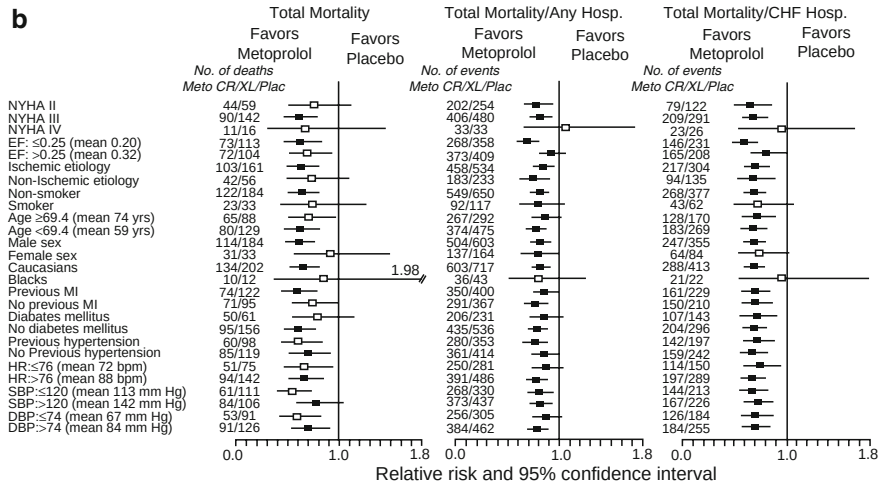
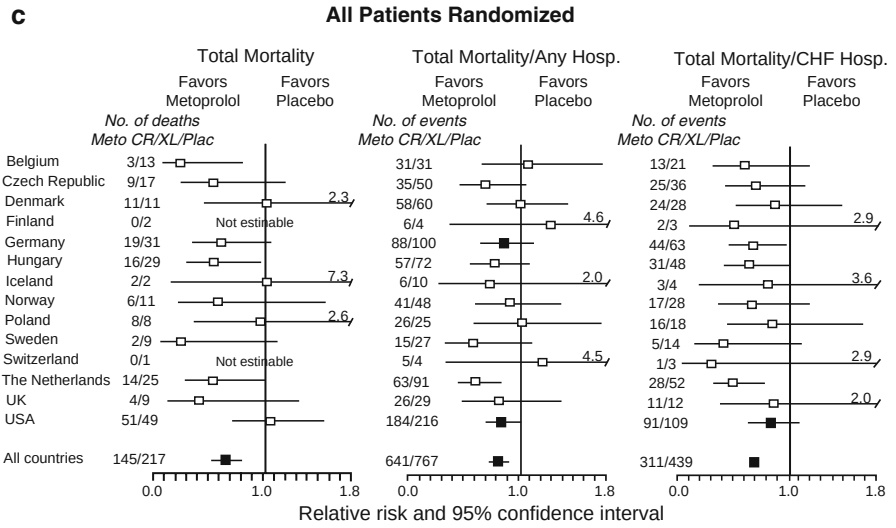


Fig. 17.6 (**a**) MERIT Kaplan–Meier estimates of cumulative percentage of total mortality after randomization – *p* value nominal and adjusted for two interim analyses (MERIT) [28]. Reproduced with the permission of Elsevier Ltd. for *Lancet*





Absolute numbers, point estimates of the hazards ratios, and 95% confidence intervals (for combined end point time to first event) for total mortality, total mortality plus all-cause hospitalization, and total mortality plus hospitalization for worsening heart failure in post hoc subgroups according to country (all patients randomized). *Filled squares*, Subgroups with a total of 180 events or more; *open squares*, subgroups with a total of < 180 events (low power).

Fig. 17.6 (continued) (**b**) Relative risk and 95% confidence intervals for selected subgroups in the MERIT trial, for total mortality, total mortality and all hospitalization, and total mortality and heart failure hospitalization [122]. Reproduced with the permission of Elsevier Ltd. for the *Am Heart J*. (**c**) Relative risk and 95% confidence intervals for the MERIT trial, for outcomes of total mortality, total mortality and hospitalization for any cause, and total mortality and heart failure hospitalization [122]. Reproduced with the permission of Elsevier Ltd. for the *Am. Heart J.*

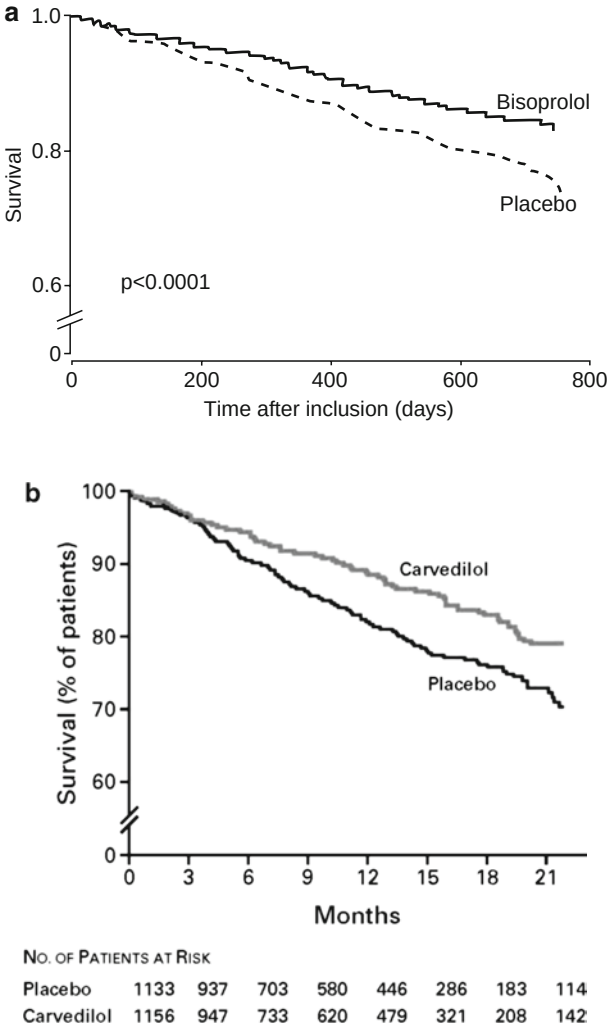


Fig. 17.7 (**a**) Kaplan–Meier survival curves for the CIBIS-II trial, comparing bisoprolol and placebo [31]. Reproduced with the permission of Elsevier Ltd. for *Lancet*. (**b**) Kaplan–Meier Analysis of Time to Death for COPERNICUS trial, comparing Placebo and Carvedilol Group. The 35% lower risk in the carvedilol group was significant: $p=0.00013$ (unadjusted) and $p=0.0014$ (adjusted) [29]

explanations for the findings, while necessary, are not sufficient. Given almost any outcome, reasonable sounding explanations can be put forward.

Often, attention is focused on subgroups with the largest intervention-control differences. However, even with only a few subgroups, the likelihood of large but spurious differences in effects of intervention between the most extreme subgroup outcomes can be considerable [111–113]. Because large, random differences

can occur, subgroup findings may easily be over-interpreted. Peto has argued that observed quantitative differences in outcome between various subgroups are to be expected, and they do not imply that the effect of intervention is truly dissimilar [107].

It has also been suggested that, unless the main overall comparison for the trial is significant, investigators should be particularly conservative in reporting significant subgroup findings [9, 111]. Lee and colleagues conducted a simulated randomized trial, in which participants were randomly allocated to two groups, although no intervention was initiated [123]. Despite the expected lack of overall difference, a subgroup was found which showed a significant difference.

In summary, subgroup analyses are important. However, they must be defined using baseline data and interpreted cautiously.

Not Counting Some Events

In prevention trials, the temptation is not to count events that are observed in the immediate postrandomization follow-up period. The rational for this practice is that events occurring that rapidly must have existed at screening, but were not detected. For example, if a cancer prevention trial randomized participants into a vitamin versus placebo trial, any immediate postrandomization cancer events could not have been prevented since the cancer had to have already been present subclinically at entry. Because the intervention could not have prevented these cases, their inclusion in the design only dilutes the results and decreases power. While such an argument has some appeal, it must be viewed with caution. Rarely are mechanisms of action of therapies or interventions fully understood. More importantly, negative impact of interventions having a more immediate effect might not be seen as easily or as quickly with this approach. If used at all, and this should be rarely, the data must be presented in both ways, i.e., with and without the excluded events.

An extreme case of dropping early events might be in a surgical or procedure trial. Participants assigned to the procedure might be at higher risk of a fatal or irreversible event. These early risks to the participant are part of the overall intervention effect and should not be eliminated from the analysis.

Some trials have defined various counting rules for events once participants have dropped out of the study or reached some level of nonadherence. For example, the Anturane Reinfarction Trial [24] suggested that no events after 7 days going off study medication should be counted. It is not clear what length of time is appropriate to eliminate events to avoid bias. For example, if a participant with an acute disease continues to decline and is removed from therapy, bias could be introduced if the therapy itself is contributing to the decline due to adverse effects and toxicity. In the APPROVe trial [71–74] described earlier in this chapter, the decision not to count events after 14 days and not to follow participants after that period of time led to controversy. In fact, the results and the interpretation were different once the almost complete follow-up was obtained [74].

Comparison of Multiple Variables

If enough significance tests are done, some of the tests may be significant by chance alone. This issue of multiple comparisons includes repeated looks at the same response variable (Chap. 15) and comparisons of multiple variables. Many clinical trials have more than one response variable, and certainly several baseline variables are measured. Thus, a number of statistical comparisons are likely to be made. These would include testing for differences in baseline characteristics as well as subgroup analyses. For example, if an investigator has 100 independent comparisons, five of them, on the average, will be significantly different by chance alone if she uses 0.05 as the level of significance. The implication of this is that the investigator should be cautious in the interpretation of results if she is making multiple comparisons. The alternative is to require a more conservative significance level. As noted earlier, lowering the significance level will reduce the power of a trial. The issue of multiple comparisons has been discussed by Miller [124], who reviewed many proposed approaches.

One way to counter the problem is to increase the sample size so that a smaller significance level can be used while maintaining the power of the trial. However, in practice, most investigators could probably not afford to enroll the number of participants required to compensate for all the possible comparisons that might be made. As an approximation, if investigators are making k comparisons, each comparison should be made at the significance level α/k, a procedure known as the Bonferonni correction [124]. Thus, for $k = 10$ and $\alpha = 0.05$, each test would need to be significant at the 0.005 level. Sample size calculations involving a significance level of 0.005 will dramatically increase the required number of participants. The Bonferonni correction is quite conservative in controlling the overall α level or false positive error rate. Therefore, it may be more reasonable to calculate sample size based on one primary response variable, limit the number of comparisons, and be cautious in claiming significant results for other comparisons.

However, there are other procedures to control the overall α level and we summarize briefly two of them [125, 126]. Assume that we prespecify m hypotheses to be tested, involving either multiple outcomes, multiple subgroups, or a combination. The goal is to control the overall α level. One implementation of the Holm procedure [125] is to order the p values from smallest to largest as $p(1), p(2), \ldots, p(m)$, corresponding to the m hypotheses $H(1), H(2), \ldots, H(m)$. Then the Holm procedure would reject $H(1)$, if $p(1) \leq \alpha/m$. If and only if $H(1)$ is rejected can we consider the next hypothesis. In that case, $H(2)$ can be rejected if $p(2) \leq \alpha/(m-1)$. This process continues until we fail to reject and then the testing must stop. The Holm procedure can also be applied if the m hypotheses can be ordered according to their importance. Here, the most important hypothesis $H(1)$ can be rejected only if the corresponding p value is less than α/m. If rejected, the next most important hypothesis $H(2)$ can be rejected if the p value is less than $\alpha/(m-1)$.

Hochberg's procedure [126] also requires that the m hypotheses be specified in advance and orders the p-values from smallest to largest as does Holm's. The Hochberg procedure allows all m hypotheses to be rejected if $p(m) \leq \alpha/m$. If this is not the case, then the remaining $m-1$ hypotheses can be rejected if $p(m-1) \leq \alpha/(m-1)$. This process is carried out for all of the m hypotheses until a rejection is obtained and then stops. Each of these procedures will not give exactly the same rejection pattern; so it is important to prespecify which one will be used.

In considering multiple outcomes or subgroups, it is important to evaluate the consistency of the results qualitatively, and not stretch formal statistical analysis too far. Most formal comparisons should be stated in advance. Beyond that, one engages in observational data analysis to generate ideas for subsequent testing.

Use of Cutpoints

Splitting continuous variables into two categories, for example by using an arbitrary "cutpoint," is often done in data analysis. This can be misleading especially if the cutpoint is suggested by the data. As an example, consider the constructed dataset in Table 17.6. Heart rate, in beats per minute, was measured prior to intervention in two groups of 25 participants each. After therapies A and B were administered, the heart rate was again measured. The average changes between groups A and B are not sufficiently different from each other ($p = 0.75$) using a standard t-test. However, if these same data are analyzed by splitting the participants into "responders" and "non-responders," according to the magnitude of heart rate reduction, the results can be made to vary. Table 17.7 shows three such possibilities, using reductions of 7, 5, and 3 beats per minute as definitions of response. As indicated, the significance levels, using a chi-square test or Fisher's exact test, change from not significant to significant and back to not significant. This created example suggests that by manipulating the cutpoint, one can observe a significance level less than 0.05 when there does not really seem to be a difference.

In an attempt to understand the mechanisms of action of an intervention, investigators frequently want to compare participants from two groups who experience the same event. Sometimes, this retrospective look can suggest factors or variables by which the participants could be subgrouped. If some subgroup is suggested, the investigator should create that subgroup in each study group and make the appropriate comparison. For example, she may find that participants in the intervention group who died were older than those in the control group who died. This retrospective observation might suggest that age is a factor in the usefulness of the intervention. The appropriate way to test this hypothesis would be to subgroup all participants by age and compare intervention versus control for each age subgroup.

Table 17.6 Differences in pre- and post-therapy heart rate, in beats per minute (HR), for groups A and B, with 25 participants each

Observation number	A			B		
	Pre HR	Post HR	Change in HR	Pre HR	Post HR	Change in HR
1	72	72	0	72	70	2
2	74	73	1	71	68	−3
3	77	71	6	75	74	1
4	73	78	−5	74	71	3
5	70	66	4	71	73	−2
6	72	76	−4	73	78	−5
7	72	72	0	71	69	2
8	78	76	2	70	74	−4
9	72	80	−8	79	78	1
10	78	71	7	71	72	−1
11	76	70	6	78	79	−1
12	73	77	−4	72	75	−3
13	77	75	2	73	72	1
14	73	79	−6	72	69	3
15	76	76	0	77	74	3
16	74	76	−2	79	75	4
17	71	69	2	77	75	2
18	72	71	1	75	75	0
19	68	72	−4	71	70	1
20	78	75	3	78	74	4
21	76	76	0	75	80	−5
22	70	63	7	71	72	−1
23	76	70	6	77	77	0
24	78	73	5	79	76	3
25	73	73	0	79	79	0
Mean	73.96	73.20	0.76	74.40	73.96	0.44
Standard deviation	2.88	3.96	4.24	3.18	3.38	2.66

Table 17.7 Comparison of change in heart rate in group *A* versus *B* by three choices of cutpoints

Beats/min	<7	≥7	<5	≥5	<3	≥3
Group *A*	25	2	19	6	17	8
Group *B*	25	0	25	0	18	7
Chi-square	$p=0.15$		$p=0.009$		$p=0.76$	
Fisher's exact	$p=0.49$		$p=0.022$		$p=0.99$	

Noninferiority Trial Analysis

As discussed in Chap. 5, noninferiority trials are challenging to design, conduct, and analyze. We pointed out the special challenges in setting the margin of noninferiority. However, once that margin of noninferiority is established prior to the start of the

trial, there remain several issues that must be included in a rigorous analysis and reported because of the clinical and regulatory implications [127–144]. If we define I to be the new intervention, C to be the control or standard, and P to be the placebo or no treatment, then we obtain from the noninferiority trial an estimate of the relative risk (RR) of I to C, RR(T/C), or an absolute difference. In the design, the metric must be established since the sample size and the interim monitoring depend on it. The first analytic challenge is to establish whether the new intervention met the criteria for noninferiority, a part of which is demonstrating that the 95% confidence interval of the estimate was less than the noninferiority margin.

As shown in Fig. 17.8, from Pocock and Ware [129], if the upper limit of the 95% confidence interval for the relative risk is less than unity, various degrees of evidence exist for superiority (See case A). For noninferiority trials, if the upper 95% confidence interval is less than the margin of noninferiority (δ), then there is evidence for noninferiority (see cases B and C). Failure to be less than this margin does not provide evidence for noninferiority (see case D). The design must have sufficient sample size and power to rule out a margin of noninferiority as discussed in Chap. 8. Although not expected when the study was designed, a noninferiority trial might also indicate harm (E).

The second desired goal of a noninferiority analysis is to demonstrate that the new intervention would have beaten a placebo or no treatment if it had been included; that is, the estimate of RR(I/P). Analytically, this can be accomplished by recognizing that RR(I/P) = RR(I/C) RR(C/P). However, for this imputation step to work requires at least two critical assumptions: (1) there is constancy of the control effect over time, and (2) the population where the control was tested against placebo is relevant to the current use where the intervention (I) is being tested. These assumptions are difficult, perhaps impossible, to establish (see Chap. 5). In this chapter, we will focus our attention on the first challenge of establishing whether or not the intervention versus control comparison was less than the noninferiority margin.

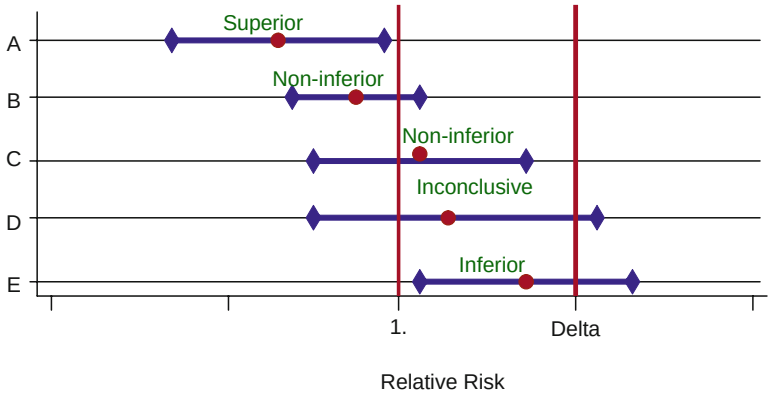


Fig. 17.8 Relative risks and 95% confidence intervals for a series of superiority and noninferiority trials modified from [129]

Assuming that an appropriate active control was selected, the trial must implement that control in a way that is consistent with best practice and is as good or better than that what was done in the initial trial that showed it to be beneficial [144]. Otherwise, the new intervention is being compared to a control that is handicapped, making it easier for the new intervention to appear similar or even better than the control. Poor adherence and conduct work in favor of the new intervention in a noninferiority trial while working against the new intervention in a superiority trial [128]. Thus, as discussed in Chap. 14, adequate measures of adherence must be collected during the trial in order to make this critical assessment. Adherence in this case does not only mean whether the participant took all or almost all of the intervention and control drugs. What else participants were taking as concomitant medication is also a consideration. If there is a substantial imbalance, interpretation of the results would be difficult.

Another key factor is whether the outcomes chosen are true measures of the effect of both the new intervention and the control. This is sometimes referred to as *assay sensitivity* [130]. Thus, whether consciously or not, an investigator might select an outcome that would show no change no matter what intervention was being studied and thus guarantee that the noninferiority margin would be achieved. Outcomes should be similar to those used in the initial control versus placebo trials.

There is a debate whether the intention-to-treat analysis or the "on treatment" analysis is most appropriate for a noninferiority designed trial. If intention-to-treat is used, nonadherence dilutes whatever differences may exist and thus is biased toward noninferiority. An "on treatment" analysis compares only those who are good adherers, or at least took some predefined portion of the intervention and thus is closer to testing the true effect. However, as we demonstrated earlier in this chapter, analyzing trials by adherence to an intervention can be substantially biased, the direction of which cannot be predicted. Thus, we do not recommend such an analysis because of the uncertainty of bias and its direction, and instead recommend that a trial be designed to minimize nonadherence. The true comparison of the new intervention may be somewhere in between the intention-to-treat and the "on treatment" but there is no dependable way to tease that estimate out. If both analytic approaches confirm noninferiority, then we can be reasonably assured of our conclusions, assuming that the noninferiority margin is reasonable [132].

Any trial relies on an adequate sample size to have power to detect differences of interest, whether this is for superiority or noninferiority. For a superiority trial, inadequate sample size works against finding differences but for noninferiority, inadequate sample size favors finding noninferiority as long as the sample size is not too small. There is a difficult balance between having a noninferiority margin that is too small and thus requiring an unachievable sample size and having a margin that is so large that the sample size is appealing but the results would not be convincing.

There are many examples of noninferiority trials, but we use one to illustrate the challenges: the Stroke Prevention using an ORal Thrombin Inhibitor in atrial Fibrillation (SPORTIF)-V trial in participants with a trial fibrillation comparing a new intervention, ximelegatran, against a standard warfarin intervention [134], with

a primary outcome of stroke incidence. A number of issues were involved. First, there were no good warfarin versus placebo trials to set the noninferiority margin. Second, the trial used absolute difference as the metric, assuming the event rate would be around 3%, but instead observed an event rate less than half of that. Thus, the noninferiority margin of 2% that was prespecified was too large given the small event rate. If the observed event rate of 1.5% had been assumed, the prespecified margin would have been much less, perhaps closer to 1%. The observed stroke rates were 1.2% in the warfarin group and 1.6% in the ximelegatran group with a 95% CI of −0.13 to 1.03% which would meet the initial margin of noninferiority. However, this was not adequate for a margin of 1%. Therefore, even though margins may be set in advance, results may invalidate the assumptions and thus the margin itself.

As discussed in Chap. 19, presentation of the results of noninferiority trials are more complex than for superiority trials because all of the assumptions must be so carefully and clearly laid out [129].

Meta-Analysis of Multiple Studies

Often in an area of clinical research, several independent trials, some of which may be large multicenter trials, are conducted over a period of a few years. Some of the trials may be too small to be conclusive on their own, but may have served as a pilot for a larger subsequent study. Investigators from a variety of medical disciplines often feel compelled to review the cumulative data on similar trials using similar participants and similar intervention strategies and try to reach a consensus conclusion of the overall results [145–153]. If this overview is performed by a formal process and with statistical methods for combining all the data with a single analysis, the analysis is usually referred to as a meta-analysis or systematic review. The methods used were essentially described in 1954 by Cochran [154] and later by Mantel and Haenszel [155]. Other authors have summarized the methodologic approaches [156–163]. The Cochrane Collaboration has been a major contributor to systematic reviews of controlled trials [164], often organized around a specific health care area or issue, including systematic reviews of adverse effects (http://aemg.cochrane.org). In addition, this group provides advice on how to conduct such systematic reviews. There are numerous examples of meta-analysis in a variety of medical disciplines and a few are referenced here [165–175]. A great deal has been written and discussed about the usefulness and challenges of meta-analyses [176–187].

Rationale and Issues

There are several reasons for conducting a systematic review or meta-analysis [145]. Probably the most common reason is to obtain more precise estimates of an intervention effect and to increase the power to observe small, but clinically important

effects. A closely related reason is to evaluate the generalizability of the results across trials, populations, and specific interventions. Subgroup analyses within a trial are often based on too few participants to be definitive or identify qualitative differences in effect. Many are also posthoc and thus unreliable due to both multiplicity of testing and data exploration. However, meta-analysis offers the opportunity to examine a limited number of prespecified hypotheses identified in individual trials. One goal of this type of subgroup meta-analysis is to guide clinicians in their practice by selecting participants most suitable for the intervention. Many submissions to the U.S. Food and Drug Administration include a meta-analysis as part of the report. If a major clinical trial is being initiated, a sensible approach is to base many aspects of the design on the summary of all existing data. This would include definitions of population and intervention, control group response rates, expected size of the intervention effect, and length of follow-up. Use of meta-analysis is a systematic process that can provide this critical information. Finally, if a new treatment or intervention gains widespread popularity early in its use, a meta-analysis may provide a balanced perspective and may suggest the need for a single, large, properly designed clinical trial to provide a more definitive result. Alternatively, meta-analyses are mandated if the opportunity to conduct a new large study no longer exists due to a loss of equipoise, even if this loss is not well justified. In this case, a meta-analysis may be the only solution to try to salvage the most reliable consensus.

As indicated, a meta-analysis is the combination of results from similar participants evaluated by similar protocols and interventions. The ideal meta-analysis is the standard analysis of a large multicenter trial, stratified by clinical center. Each center plays the role of a small study. Protocols and treatment strategies are identical, and participants are more similar compared to a typical collection of trials. Meta-analysis should never be an excuse for conducting many small studies, loosely connected with the expectation that meta-analysis will rescue the definitive result.

The concept that the ideal meta-analysis is the large multicenter trial focuses on some of the limitations of the typical meta-analysis. While differences exist in the implementation of a clinical protocol across centers, these differences are less than for a collection of independently conducted large or small trials. Typically in meta-analysis, nontrivial differences exist in actual treatment, study population, length of follow-up, measures of outcome, and quality of data [177–182]. With these and other potential differences, the decision as to which studies are similar enough to justify combining their data represents a challenge.

Many support the concept that the most valid overview and meta-analysis requires all studies conducted be available for inclusion or at least for consideration [145, 178]. Furberg [182] provides a review of seven meta-analysis of lipid lowering trials. Each article presents different inclusion criteria, such as the number of participants or the degree of cholesterol reduction. The results vary depending on the criteria used. As in a clinical trial protocol, the questions and the criteria should be stated in advance [184]. While it is already difficult to decide what similar might mean, a further serious complication is that all trials conducted may not be readily accessible in the literature due to publication bias [183, 184]. Furthermore, not all

trials that are started are completed or published. The problem is that trials published are more likely to be positive ($p<0.05$) or favorable. Trials that yield neutral or indifferent results are less likely to be published. One example described by Furberg and Morgan [177] illustrates the problem. One overview [183] of the use of propranolol in patients following a heart attack reported 7 of 45 patients died in the hospital compared to a nonrandomized, placebo-control where 17 of 46 died. Controversy over design limitations caused the investigator to conduct two additional randomized trials, but neither were ever published. One showed no difference and the other a negative (harmful) trend. As a further complication, Chalmers et al. [185] pointed out that a MEDLINE literature search may only find 30–60% of published trials. This is due in part to the way results are presented and searches of typical key words may not uncover relevant papers. Although search engines may be better now, there are undoubtedly still limitations.

Another bias, referred to as investigator bias, is that what outcome variables get reported may depend on the investigator. If protocols were adhered to strictly, investigator bias may not be a problem. However, repeated testing, multiple subgroups, and multiple outcomes may not be easy to detect from the published report [176]. Early promising results may draw major attention, but if later results show smaller intervention effects, they may go unnoticed or be harder to find for the systematic review. Furthermore, authors of overviews are also to subject to investigator bias. That is, unless the goals of the meta-analysis are clearly stated a priori, a positive result can be found in this analysis by sifting through numerous attempts. In fact, data dredging for large studies is more likely to find at least one positive result than for a single small study. A great deal of time and persistence are required in order to get access to all known conducted trials and accurately extract the relevant data. Not all meta-analyses are conducted with the same degree of thoroughness.

The medical literature is filled with meta-analysis of trials covering a wide range of disciplines [165–175]. Chalmers and colleagues [167] reviewed six small studies that used anticoagulants in an effort to reduce mortality in heart attack patients. While only one of the six was individually significant, the combined overall results suggested a statistically significant 4.2% absolute reduction in mortality. The authors suggested no further trials were necessary. However, due to issues raised, this analysis drew serious criticism [176]. Several years later, Yusuf and colleagues [174] reviewed 33 fibrinolytic trials, focusing largely on the use of streptokinase. This overview included trials with many dissimilarities in dose, route and time of administration, and setting. Although the meta-analysis for intravenous use of fibrinolytic drugs was impressive, and the authors felt that results were not due to reporting biases, they nevertheless discussed the need for future large-scale trials before widespread use should be recommended. There were issues, for example, as to how quickly such an intervention needed to be started after onset of a heart attack. Thus, timing needed to be resolved. Canner [170] conducted an overview of six randomized clinical trials testing aspirin use in participants with a previous heart attack to reduce mortality. His overall meta-analysis suggested a 10% reduction although not significant ($p=0.11$). However, there was an apparent heterogeneity

of results and the largest trial had a slightly negative mortality result. The Canner overview was repeated by Hennekens et al. [175] after several more trials had been conducted. This updated analysis demonstrated favorable results.

May et al. [168] conducted an early overview of several modes of therapy for secondary prevention of mortality after a heart attack. Their overview covered anti-arrhythmic drugs, lipid-lowering drugs, anticoagulant drugs, beta-blocker drugs, and physical exercise. Although statistical methods were available to combine studies within each treatment class, they chose not to combine results, but simply provided relative risks and confidence interval results graphically for each study. A visual inspection of the trends and variation in trial results suggests a summary analysis. Yusuf et al. [173] later provided a more detailed overview of beta blockade studies. While using a similar graphical presentation, they calculated a summary odds ratio and its confidence interval. The details of the method are described below. Meta-analysis of cancer trials have also been conducted including the use of adjuvant therapy for breast cancer [172]. While using multiple chemotherapeutic agents indicated improved relapse-free survival after 3 and 5 years of follow-up, as well as for survival, the dissimilarity among the trials led the authors to call for more trials and better data.

Thompson [186] pointed out the need to investigate thoroughly sources of heterogeneity such as clinical differences across studies. These differences may be in populations studied, intervention strategies, outcomes measured, or other logistical aspects. Given such differences, incompatible results among individual studies might be expected. Statistical tests for heterogeneity often have low statistical power even in the presence of a moderate heterogeneity. Thompson [186] argued that we should investigate the influence of apparent clinical differences between studies and not rely on formal statistical tests to give us assurance of no heterogeneity. In the presence of apparent heterogeneity, overall summary results should be interpreted cautiously. Thompson described an example of a meta-analysis of 28 studies evaluating cholesterol lowering and the impact on risk of coronary heart disease. A great deal of heterogeneity was present, so a simple overall estimate of risk reduction may be misleading. He showed that factors such as age of the cohort, length of treatment, and size of study were contributing factors. Taking these factors into account made the heterogeneity less extreme and results more interpretable. One analysis showed that the percent reduction in risk decreased with the age of the participant at the time of the event, a point not seen in the overall meta-analysis. However, he also cautioned that such analyses of heterogeneity must be interpreted cautiously, just as for subgroup analyses in any single trial.

Meta-analysis, as opposed to typical literature reviews, usually puts a p value on the conclusion. The statistical procedure may allow for calculation of a p value, but it implies a precision which may be inappropriate. The possibility that studies may be missed and the issue of study selection may make the interpretation of the p value tenuous. As indicated, quality of data may vary from study to study. Data from some trials may be incomplete, and perhaps not even recognized as such. Thus, only very simple and unambiguous outcome variables, such as all-cause mortality and major morbid events ought to be used for meta-analysis.

Statistical Methods

Since the meta-analysis became a popular approach to summarizing a collection of studies, numerous statistical publications have been produced addressing several technical aspects [153–163]. Most of this is beyond the technical scope of this text. However, we will summarize one popular meta-analytic method that combines information on success and failure by study group across separate trials.

A standard approach as described in the overview paper by May et al. [168] is to summarize each study with an odds ratio, or a relative risk, along with a 95% confidence interval. That is, suppose each trial can be summarized by a 2×2 table where S represents success and F represents failure, and a, b, c, d are the numbers of individuals in each category.

	Result		
Group	S	F	TOTAL
Treatment	a	b	$a+b$
Control	c	d	$c+d$
TOTAL			m

Each study compares the success rate in the intervention arm (P_I) and control arm (P_C). Using this table, the estimate for $P_I = a/(a+b)$ and the estimate for $P_C = c/(c+d)$. The relative risk $RR = P_I/P_C$ is one summary statistic. The estimate for RR is $a(c+d)/b(c+d)$. Another summary statistics, the odds ratio (OR), that approximates the RR, may also be used. An estimate of the OR is ad/bc and the 95% confidence interval is

$$\frac{ad}{bc}\exp\left[\pm 1.96\sqrt{1/a+1/b+1/c+1/d}\right]$$

Typically, the OR estimate and 95% confidence interval are plotted in a single graph for each trial to provide a visual summary. This may be seen in May et al. [168] or in Yusuf et al. [173]. Figure 17.9, from Yusuf et al. [174], summarizes the effects of 24 trials of fibrinolytic treatment on mortality in people with an acute heart attack. The hash mark represents the estimated OR, and the line represents the 95% confidence interval. May et al. [168] went no further in their overview. Yusuf et al. [174], however, recommended that a single estimation of the OR be obtained, combining all studies.

Two technical approaches are used for this situation, both suggested by Cochran [154] in 1954. If all trials included in the meta-analysis are estimating the same true (but unknown) fixed effect of an intervention, the Mantel–Haenszel method [155] is used with a slight variation. This is similar to the logrank or Mantel–Haenszel method in the chapter on survival analysis. This method is referred to here as the Peto–Yusuf method. If the trials are assumed to have dissimilar or heterogeneous true intervention effects, the effects are described by a random effects model, a method suggested by DerSimonian and Laird [160].

The Peto–Yusuf [173] method follows the Cochran–Mantel–Haenszel procedure [154, 155]. Let $O_i = a_i$, E_i be the expected value of O_i and V_i be the variance. Then,

$$O_i = a_i$$

$$E_i = \frac{(a_i + c_i)(a_i + b_i)}{m_i}$$

$$V_i = \frac{(a_i + c_i)(b_i + d_i)(a_i + b_i)(c_i + d_i)}{m_i^2 (m_i - 1)}$$

and let $O = \sum_i O_i$, $E = \sum_i E_i$, and $V = \sum V_i$, where $i = 1, 2, 3, \ldots, N$. As shown in Chap. 15, the statistic $Z_{\text{MH}} = (O - E)/\sqrt{V}$ has a standard normal distribution. The Peto–Yusuf [173] method estimates the pooled odds ratio, OR_p, as

$$\text{OR}_\text{p} = \exp\left[(O - E)/V\right]$$

and the 95% confidence interval as

$$\exp\left\{\frac{O - E}{V} \pm 1.96/\sqrt{V}\right\}$$

Using this method, the summary pooled odds ratio and 95% confidence interval, shown in Fig. 17.9, can be computed for the 24 fibrinolytic studies. The overall estimate of the pooled studies is shown in the last line. The size of the symbol in these plots, sometimes referred to as "forest plots," is an indication of the size of each individual studies.

The method of DerSimonian and Laird [160] compares rate differences within each study, and obtains a pooled estimate of the rate difference as well as the standard error. The pooled estimate of the rate difference is a weighted average of the individual study rate differences. The weights are the inverse of the sum of the between and within study variance components of intervention effect. If the studies are relatively similar or homogeneous in intervention effect, the two methods provide very similar results [159]. Heterogeneity tests generally are not as powerful as the test for main effects. However, if studies vary in intervention effect, these two methods can produce difference results as illustrated by Berlin et al. [159] as well as Pocock and Hughes [157].

In the presence of serious heterogeneity of treatment effect, the appropriateness of obtaining a single point estimate must be questioned. This was part of the rationale for May et al. [168] not combining studies. If the heterogeneity is qualitative, that is, some estimates of the OR are larger than unity and others less than unity, then a combined single estimate is perhaps not wise. This would be especially true if these estimates indicated a time trend, which could occur if dose and participant selection changed as more experience with the new intervention was obtained.

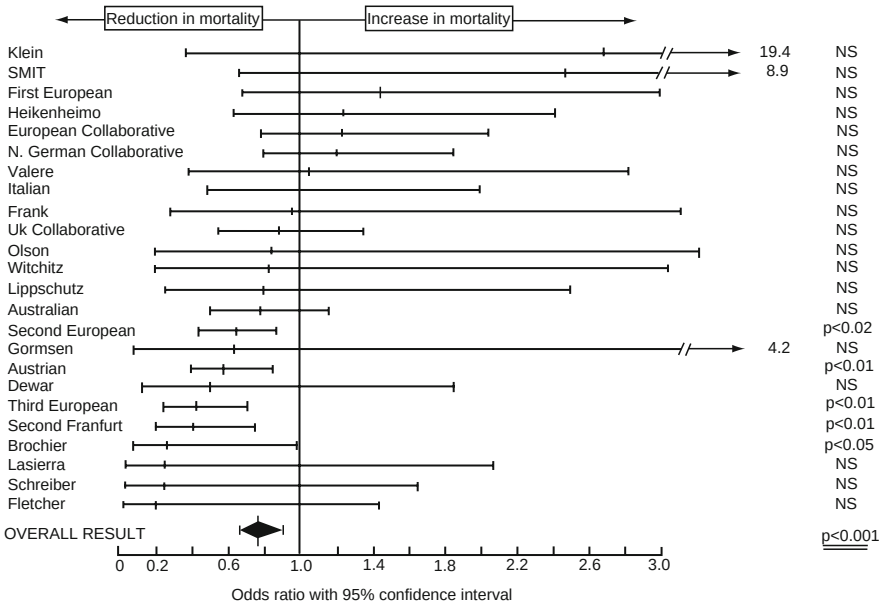


Fig. 17.9 Apparent effects of fibronolytic treatment on mortality in the randomized trials of IV treatment of acute myocardial infarction. (Reproduced with permission of the Editor, *European Heart Journal* and Dr. S. Yusuf [174])

Whether a fixed effects or a random effects model is preferable is a matter of debate, but neither are exactly correct. The random effects model has an undesirable aspect, in that small trials may dominate the final estimate whereas with the fixed effect model, larger trials get greater weight. However, the meta-analysis is conducted on available trials, none of which are typically very representative of the general population to which the intervention may be applied. That is, the trials that are available do not contain a random sample of people from the targeted population but rather are participants who volunteered and who in other respects may not be representative. Thus, the estimate of the intervention effect is not as relevant as whether or not the intervention is at all effective. We prefer a fixed affects model but suggest that both models should be conducted to examine what, if any, differences exist.

Chalmers, a strong advocate of clinical trials, argued that participants should be randomized early in the evolution and evaluation of a new intervention [188]. Both as a result of that kind of advocacy and the fact that small trials are always done before large ones in the development of new interventions, an early meta-analysis is likely to consist of many small studies. Sometimes, meta-analyses of just small trials might yield significant results.

Thus, meta-analyses are seen by many as alternatives to the extraordinary effort and cost often required to conduct adequately powered individual trials. Rather than providing a solution, they perhaps ought to be viewed as a way of summarizing

existing data; a way that has strengths and weaknesses, and must be critically evaluated. It would clearly be preferable to combine resources prospectively and collaborate in a single large study. Pooled studies cannot replace individual, well-conducted multicenter trials.

Analysis Following Trend Adaptive Designs

As discussed in Chaps. 5 and 16, the design of a trial may have an adaptive element. This might be a group sequential design for early termination due to overwhelming benefit or a strong signal for harm, or perhaps futility. Among the adaptive designs were those that altered the sample size. Some of these sample size changes are due to overall lower event rates or higher variability in the primary outcome than was assumed in the original sample size estimate. In these instances, the final analysis proceeds as normal. However, another method for sample size change relies on trend adaptive designs. In these designs, which depend on the emerging trend in the data, the final critical value or significance level will be affected and thus must be kept in mind for the final analysis.

For example, some trials may monitor accumulating interim data and may terminate the trial early for evidence of benefit or harm. If a group sequential design using a 0.05 two-sided significance level O'Brien-Fleming boundary were used five times during the trial, approximately equally spaced, the final critical value would not be +1.96 and −1.96 for the upper and lower bounds but a value closer to 2.04.

For trend adaptive sample size changes, the final critical value depends on which methodology was used but all will require typically a more conservative value, for example, than a two-sided nominal alpha level of 0.05 (a critical value of 1.96).

Other than adjusting the final critical value, the analyses for these trend adaptive designs may also utilize a modified test statistic. For example, if the method of Cui et al. [189] is used in increasing the sample size, a weighted test statistic as described in Chap. 16 is required. Future observations are given less weight that the early existing observations. The usual test statistic is not appropriate in this situation. For the other trend adaptive methods described in Chaps. 5 and 16, the final analysis can proceed with the standard statistics in a usual straightforward fashion, adjusting for the final critical value from sequential testing as appropriate.

Appendix

Mantel–Haenszel Statistic

Suppose an investigator is comparing response rates and divides the data into a number of strata using baseline characteristics. For each stratum i, a 2×2 table is constructed.

2×2 Table for *i*th Stratum

	Response		
	Yes	No	
Intervention	a_i	b_i	$a_i + b_i$
Control	c_i	d_i	$c_i + d_i$
Total	$a_i + c_i$	$b_i + d_i$	n_i

The entries a_i, b_i, c_i, and d_i represent the counts in the four cells and n_i is the number of participants in the *i*th stratum. The marginals represent totals in the various categories. The value $(a_i + c_i)/n_i$ represents the overall response rate for the *i*th stratum. Within the *i*th stratum, the rates $a_i/(a_i + b_i)$ with $c_i/(c_i + d_i)$ are compared. The standard chi-square test for 2×2 tables could be used to compare group differences in this stratum. However, the investigator is interested in "averaging" the comparison over all the strata. The method for combining several 2×2 tables over all tables or strata was described by Cochran [157] and Mantel and Haenszel [158]. The summary statistic, denoted MH, is given by:

$$\mathrm{MH} = \frac{\left\{\sum_{i=1}^{K}\left[a_i - (a + c_i)(a_i + b_i)/n_i\right]\right\}^2}{\sum_{i=1}^{K}(a_i + c_i)(b_i + d_i)(a_i + b_i)(c_i + d_i)/n_i^2(n_i - 1)}$$

The MH statistic has a chi-square distribution with one degree of freedom. The square root of MH has a normal distribution. Tables for this distribution are available in standard statistical textbooks. Any value for MH greater than 3.84 is significant at the 0.05 level, and any value greater than 6.63 is significant at the 0.01 level. This method is particularly appropriate for covariates that are discrete or continuous covariates that have been classified into intervals.

References

1. Pagano M, Gauvreau K. *Principles of Biostatistics* (2nd edition). Pacific Grove, CA: Duxbury Press, 2000.
2. Piantadosi S. *Clinical Trials: A Methodologic Perspective* (2nd edition). Wiley Series in Probability & Statistics. New Jersey: John Wiley and Sons, Inc., 2005.
3. Armitage P. *Statistical Methods in Medical Research*. New York: John Wiley and Sons, 1977.
4. Hill AB. *Principles of Medical Statistics* (9th edition). New York: Oxford University Press, 1971.
5. Cook T, DeMets DL. *Introduction to Statistical Methods for Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC; Taylor & Francis Group, LLC, 2008.
6. Geller N. *Advances in Clinical Biostatistics*. New York: Marcel Dekker, 2004.
7. Woolson R. *Statistical Methods for the Analysis of Biomedical Data*. New York: John Wiley and Sons, 1987.
8. Fisher L, Van Belle G. *Biostatistics – A Methodology for the Health Sciences*. New York: John Wiley and Sons, 1993.

9. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br J Cancer* 1976;34: 585–612.
10. Armitage P. The analysis of data from clinical trials. *Statistician* 1980;28:171–183.
11. Newcombe RG. Explanatory and pragmatic estimates of the treatment effect when deviations from allocated treatment occur. *Stat Med* 1988;7:1179–1186.
12. Fleiss JL. Analysis of data from multiclinic trials. *Control Clin Trials* 1986;7:267–275.
13. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. *J Chronic Dis* 1967;20:637–648.
14. Sackett DL, Gent M. Controversy in counting and attributing events in clinical trials. *N Engl J Med* 1979;301:1410–1412.
15. ICH Expert Working Group. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. *Stat Med* 1999;18:1903–1942.
16. FDA Guidance. International Conference on Harmonization: Guidance on statistical principles for clinical trials. http://www.fda.gov/downloads/RegulatoryInformation/Guidances/UCM129505.pdf
17. May GS, DeMets DL, Friedman LM, et al. The randomized clinical trial: bias in analysis. *Circulation* 1981;64:669–673.
18. Ingle JN, Ahmann DL, Green SJ, et al. Randomized clinical trial of diethylstilbestrol versus tamoxifen in postmenopausal women with advanced breast cancer. *N Engl J Med* 1981;304:16–21.
19. The Canadian Cooperative Study Group. A randomized trial of aspirin and sulfinpyrazone in threatened stroke. *N Engl J Med* 1978;299:53–59.
20. Beta-blocker Heart Attack Trial Research Group. A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. *JAMA* 1982;247: 1707–1714.
21. The Coronary Drug Project Research Group. Clofibrate and niacin in coronary heart disease. *JAMA* 1975;231:360–381.
22. Roberts R, Croft C, Gold HK, et al. Effect of propranolol on myocardial infarct size in a randomized blinded multicenter trial. *N Engl J Med* 1984;311:218–225.
23. Collaborative Group on Antenatal Steroid Therapy. Effect of antenatal dexamethasone administration on the prevention of respiratory distress syndrome. *Am J Obstet Gynecol* 1981;141:276–287.
24. The Anturane Reinfarction Trial Research Group. Sulfinpyrazone in the prevention of sudden death after myocardial infarction. *N Engl J Med* 1980;302:250–256.
25. Temple R, Pledger GW. The FDA's critique of the Anturane Reinfarction Trial. *N Engl J Med* 1980;303:1488–1492.
26. Anturane Reinfarction Trial Policy Committee. The Anturane Reinfarction Trial: reevaluation of outcome. *N Engl J Med* 1982;306:1005–1008.
27. Soran A, Nesbitt L, Mamounas EP, et al. Centralized medical monitoring in Phase III trials: the National Surgical Adjuvant Breast and Bowel Project (NSABP) experience. *Clin Trials* 2006;3:478–485
28. MERIT-HF Study Group. Effect of metoprolol CR/XL in chronic heart failure: Metoprolol CR/XL randomised intervention trial in congestive heart failure (MERIT-HF). *Lancet* 1999;353:2001–2007.
29. Packer M, Coats AJS, Fowler MB, et al. for the Carvedilol Prospective Randomized Cumulative Survival (COPERNICUS) Study Group. Effect of Carvedilol on survival in severe chronic heart failure. *N Engl J Med* 2001;344:1651–1658.
30. Kjekshus J, Apetrei E, Barrios V, et al. for the CORONA Group. Rosuvastatin in older patients with systolic heart failure. *N Engl J Med* 2007;357:2248–2261.
31. CIBIS II Investigators and Committees. The Cardiac Insufficiency Bisoprolol Study II (CIBIS II): a randomised trial. *Lancet* 1999;353:9–13.
32. The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Engl J Med* 1993;329:673–682.

33. The Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project. *N Engl J Med* 1980;303:1038–1041.
34. The Coronary Drug Project Research Group. Initial findings leading to modifications of its research protocol. *JAMA* 1970;214:1303–1313.
35. Verter J, Friedman L. Adherence measures in the Aspirin Myocardial Infarction Study (AMIS) (abstract). *Control Clin Trials* 1984;5:306.
36. Wilcox RG, Roland JM, Banks DC, et al. Randomised trial comparing propranolol with atenolol in immediate treatment of suspected myocardial infarction. *Br Med J* 1980;280:885–888.
37. Detre K, Peduzzi P. The problem of attributing deaths of nonadherers: the VA coronary bypass experience. *Control Clin Trials* 1982;3:355–364.
38. Lipid Research Clinics Program. The Lipid Research Clinics Coronary Primary Prevention Trial results. 1. Reduction in incidence of coronary heart disease. *JAMA* 1984;251:351–364.
39. Diggle PJ. Testing for random dropouts in repeated measurement data. *Biometrics* 1989;45:1255–1258.
40. Dolin R, Reichman RC, Madore HP, et al. A controlled trial of amantadine and rimantadine in the prophylaxis of influenza A infection. *N Engl J Med* 1982;307:580–584.
41. Heyting A, Tolboom JTBM, Essers JGA. Statistical handling of drop-outs in longitudinal clinical trials. *Stat Med* 1992;11:2043–2061.
42. Hoover DR, Munoz A, Carey V, and the Multicenter AIDS Cohort Study. Using events from dropouts in nonparametric survival function estimation with application to incubation of AIDS. *J Am Stat Assoc* 1993;88:37–43.
43. Lagakos SW, Lim LL-Y, Robins JM. Adjusting for early treatment termination in comparative clinical trials. *Stat Med* 1990;9:1417–1424.
44. Morgan TM. Analysis of duration of response: a problem of oncology trials. *Control Clin Trials* 1988;9:11–18.
45. Oakes D, Moss AJ, Fleiss JL, et al. and the Multicenter Diltiazem Post-Infarction Trial Research Group. Use of compliance measures in an analysis of the effect of diltiazem on mortality and reinfarction after myocardial infarction. *J Am Stat Assoc* 1993;88:44–49.
46. Pledger GW. Basic statistics: importance of compliance. *J Clin Res Pharmacoepidemiol* 1992;6:77–81.
47. Redmond C, Fisher B, Wieand HS. The methodologic dilemma in retrospectively correlating the amount of chemotherapy received in adjuvant therapy protocols with disease-free survival. *Cancer Treat Rep* 1983;67:519–526.
48. Ridout MS. Testing for random dropouts in repeated measurement data. *Biometrics* 1991;47:1617–1621.
49. Simon R, Makuch RW. A non-parametric graphical representation of the relationship between survival and the occurrence of an event: application to responder versus non-responder bias. *Stat Med* 1984;3:35–44.
50. Sommer A, Zeger SL. On estimating efficacy from clinical trials. *Stat Med* 1991; 10:45–52.
51. Pizzo PA, Robichaud KJ, Edwards BK, et al. Oral antibiotic prophylaxis in patients with cancer: a double-blind randomized placebo-controlled trial. *J Pediatr* 1983;102:125–133.
52. Dillman RO, Seagren SL, Propert KJ, et al. A randomized trial of induction chemotherapy plus high-dose radiation versus radiation alone in stage III non-small-cell lung cancer. *N Engl J Med* 1990;323:940–945.
53. The Intermittent Positive Pressure Breathing Trial Group. Intermittent positive pressure breathing therapy of chronic obstructive pulmonary disease – A clinical trial. *Ann Intern Med* 1983;99:612–620.
54. Espeland MA, Byington RP, Hire D, et al. Analysis strategies for serial multivariate ultrasonographic data that are incomplete. *Stat Med* 1992;11:1041–1056.
55. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 1987.

56. Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis.* New York: Wiley, 2004 (Chapter 13).
57. Conaway MR, Rejeski WJ, Miller ME. Statistical issues in measuring adherence: methods for incomplete longitudinal data. In Shumaker SA, Ockene JK, Riekert KA (eds.). *The Handbook of Health Behavior Change*. New York: Springer Publishing Co., 2009, pp. 375–391.
58. Rubin D. Inference and missing data. *Biometrika* 1976;63:581–592.
59. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Stat Methodol* 1977;39:1–38.
60. Efron B. Missing data, imputation, and the bootstrap. *J Am Stat Assoc* 1994;89:463–474.
61. Greenlees JS, Reece WS, Zieschang KD. Imputation of missing values when the probability of response depends on the variable being imputed. *J Am Stat Assoc* 1982;77:251–261.
62. Laird NM. Missing data in longitudinal studies. *Stat Med* 1988;7:305–315.
63. Little RJ. Modeling the drop out mechanism in repeated-measures studies. *J Am Stat Assoc* 1995;90:1112–1121.
64. Shao J, Zhong B. Last observation carry-forward and last observation analysis. *Stat Med* 2003;22:2429–2441.
65. Molenberghs G, Kenward MG. *Missing Data in Clinical Studies*. New York: John Wiley and Sons, 2007.
66. Steering Committee of the Physicians' Health Study Research Group. Final report on the aspirin component of the ongoing Physicians' Health Study. *N Engl J Med* 1989;321:129–135.
67. Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in health postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *JAMA* 2002;288:321–333.
68. Wu MC, Carrell RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 1988;44:175–188.
69. Wu MC, Bailey KR. Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics* 1989;45:939–955.
70. Bristow MR, Saxon LA, Boehmer et al., for the Comparison of Medical Therapy, Pacing, and Defibrillation in Heart Failure (COMPANION) Investigators. Cardiac-resynchronization therapy with or without an implantable defibrillator in advanced chronic heart failure. *N Engl J Med* 2004;350:2140–2150.
71. Bresalier RS, Sandler RS, Quan H, et al., for the Adenomatous Polyp Prevention on Vioxx (APPROVe) Trial Investigators. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. *N Engl J Med* 2005;352:1092–1102.
72. Lagakos S. Time-to-event analysis for long-term treatments – the APPROVe Trial. *N Engl J Med* 2006;355:113–117.
73. Nissen S. Adverse cardiovascular effects of rofecoxib. *N Engl J Med* 2006;355:203–204.
74. Baron JA, Sandler RS, Bresalier RS, et al. Cardiovascular events associated with rofecoxib: final analysis of the APPROVe trial. *Lancet* 2008;372:1756–1764.
75. Kruskal WH. Some remarks on wild observations. *Technometrics* 1960;2:1–3.
76. Dixon WJ. Processing data for outliers. *Biometrics* 1953;9:74–89.
77. Grubbs FE. Procedures for detecting outlying observations in samples. *Technometrics* 1969;11:1–21.
78. Canner PL, Huang YB, Meinert CL. On the detection of outlier clinics in medical and surgical trials: I. Practical considerations. *Control Clin Trials* 1981;2:231–240.
79. Canner PL, Huang YB, Meinert CL. On the detection of outlier clinics in medical and surgical trials: II. Theoretical considerations. *Control Clin Trials* 1981;2:241–252.
80. The Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). *Lancet* 1994;344:1383–1389.
81. Shepherd J, Cobbe SM, Ford I, et al., for the West of Scotland Coronary Prevention Study Group. Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. *N Engl J Med* 1995;333:1301–1307.

82. Anand IS, Carson P, Galle E, et al. Cardiac resynchronization therapy reduces the risk of hospitalizations in patients with advanced heart failure: results from the Comparison of Medical Therapy, Pacing and Defibrillation in Heart Failure (COMPANION) trial. *Circulation* 2009;119:969–977.
83. Cannon CP, Braunwald E, McCabe CH, et al. Intensive versus moderate lipid lowering with statins after acute coronary syndromes. *N Engl J Med* 2004;350:1495–1504.
84. Ferreira-Gonzales I, Busse JW, Heels-Ansdell D, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomized controlled trials. *BMJ* 2007;334:756–757
85. Tomlinson G, Detsky A. Composite end points in randomized trials; there is no free lunch. *JAMA* 2010;303:267–268.
86. Weiss GB, Bunce H III, Hokanson JA. Comparing survival of responders and nonresponders after treatment: a potential source of confusion interpreting cancer clinical trials. *Control Clin Trials* 1983;4:43–52.
87. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. *J Clin Oncol* 1983;1:710–719.
88. Cox DR. Regression models and life-tables. *J R Stat Soc Series B Stat Methodol* 1972;34:187–220.
89. Efron B, Feldman D. Compliance as an explanatory variable in clinical trials. *J Am Stat Assoc* 1991;86:9–17.
90. Egger MJ, Coleman ML, Ward JR, et al. Uses and abuses of analysis of covariance in clinical trials. *Control Clin Trials* 1985;6:12–24.
91. Oye RK, Shapiro MF. Reporting results from chemotherapy trials. *JAMA* 1984;252:2722–2725.
92. Rosenbaum PR. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J R Stat Soc Ser A* 1984;147:656–666.
93. Albert JM, DeMets DL. On a model-based approach to estimating efficacy in clinical trials. *Stat Med* 1994;13:2323–2335.
94. Beach ML, Meier P. Choosing covariates in the analysis of clinical trials. *Control Clin Trials* 1989;10:161S–175S.
95. Byar DP. Assessing apparent treatment – covariate interactions in randomized clinical trials. *Stat Med* 1985;4:255–263.
96. Canner PL. Covariate adjustment of treatment effects in clinical trials. *Control Clin Trials* 1991;12:359–366.
97. Canner PL. Further aspects of data analysis. *Control Clin Trials* 1983;4:485–503.
98. Crager MR. Analysis of covariance in parallel-group clinical trials with pretreatment baselines. *Biometrics* 1987;43:895–901.
99. Morgan TM, Elashoff RM. Effect of covariate measurement error in randomized clinical trials. *Stat Med* 1987;6:31–41.
100. Thall PF, Lachin JM. Assessment of stratum-covariate interactions in Cox's proportional hazards regression model. *Stat Med* 1986;5:73–83.
101. Shuster J, van Eys J. Interaction between prognostic factors and treatment. *Control Clin Trials* 1983;4:209–214.
102. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patients subsets. *Biometrics* 1985;41:361–372.
103. Aspirin Myocardial Infarction Study Research Group. A randomized, controlled trial of aspirin in persons recovered from myocardial infarction. *JAMA* 1980;243:661–669.
104. Yates F. The analysis of multiple classifications with unequal numbers in the different classes. *J Am Stat Assoc* 1934;29:51–66.
105. Report from the Committee of Principal Investigators. A co-operative trial in the primary prevention of ischaemic heart disease using clofibrate. *Br Heart J* 1978;40:1069–1118.
106. Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Commun Stat Theory Methods* 1991;20:2609–2631.

107. Peto R. Statistical aspects of cancer trials. In Halnan KE (ed.). *Treatment of Cancer*. London: Chapman and Hall, 1982.
108. Multicentre International Study. Improvement in prognosis of myocardial infarction by long-term beta-adrenoreceptor blockade using practolol. *Br Med J* 1975;3:735–740.
109. Andersen MP, Bechsgaard P, Frederiksen J, et al. Effect of alprenolol on mortality among patients with definite or suspected acute myocardial infarction. *Lancet* 1979;ii:865–868.
110. Furberg CD, Hawkins CM, Lichstein E. Effect of propranolol in postinfarction patients with mechanical or electrical complications. *Circulation* 1984;69:761–765.
111. Simon R. Patient subsets and variation in therapeutic efficacy. *Br J Clin Pharmacol* 1982;14:473–482.
112. Ingelfinger JA, Mosteller F, Thibodeau LA, Ware JH. *Biostatistics in Clinical Medicine*. New York: MacMillan, 1983, pp. 255–258.
113. Furberg CD, Byington RP. What do subgroup analyses reveal about differential response to beta-blocker therapy? The Beta-blocker Heart Attack Trial experience. *Circulation* 1983;67(suppl 1):I-98–I-101.
114. ISIS-2 (Second International Study of Infarct Survival) Study Collaborative Group. Randomized trial or streptokinase, oral aspirin, or both, or neither among 17,187 suspected cases of acute myocardial infarction ISIS-2. *Lancet* 1988;ii:349–360
115. Packer M, O'Conner CM, Ghali JK, et al. Effect of amlodipine on morbidity and mortality in severe chronic heart failure. Prospective Randomized Amlodipine Survival Evaluation Study Group. *N Engl J Med* 1996;335:1107–1114.
116. Thackray S, Witte K, Clark AL, Cleland JG. Clinical trials update: OPTIME-CHF, PRAISE-2, ALLHAT. *Eur J Heart Fail* 2000;2:209–212.
117. Helgason T, Jonasson MR. Evidence for a food additive as a cause of ketosis-prone diabetes. *Lancet* 1981;ii:716–720.
118. Dijkstra BK. Origin of carcinoma of the bronchus. *J Natl Cancer Inst* 1963;31:511–519.
119. Davies JM. Cancer and date of birth. *Br Med J* 1963;ii:1535.
120. Bass C, Strackee J, Jones I. Lung cancer and month of birth (Letter). *Lancet* 1964;i:47.
121. Goudie RB. The birthday fallacy and statistics of Icelandic diabetes (Letter). *Lancet* 1981;ii:1173.
122. Wedel H, DeMets D, Deedwania P, et al., on behalf of the MERIT-HF Study Group. Challenges of subgroup analyses in multinational clinical trials: experiences from the MERIT-HF trial. *Am Heart J* 2001;142:502–511.
123. Lee KL, McNeer JF, Starmer CF, et al. Clinical judgment and statistics. Lessons from a simulated randomized trial in coronary artery disease. *Circulation* 1980;61:508–515.
124. Miller RG Jr. *Simultaneous Statistical Inference*. New York: McGraw-Hill, 1966.
125. Holm S. A simple sequentially rejective multiple test procedure. *Scand Stat Theory Appl* 1979;6:65–70.
126. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988;75:800–802.
127. Piaggio G, Elbourne DR, Altman DG, for the CONSORT Group. Reporting of noninferiority and equivalence randomized trials. An extension of the CONSORT statement. *JAMA* 2006;295:1152–1160.
128. Koch A, Rohmel J. The impact of sloppy conduct of noninferiority studies. *Drug Info J* 2002;36:3–6.
129. Pocock SJ, Ware JH. Translating statistical findings into plain English. *Lancet* 2009;373:1926–1928.
130. Kaul S, Diamond GA. Good enough: a primer on the analysis and interpretation of noninferiority trials. *Ann Intern Med* 2006;145:62–69.
131. Diamond GA, Kaul S. An Orwellian discourse on the meaning and measurement of noninferiority. *Am J Cardiol* 2006;99:284–287.
132. Kaul S, Diamond GA. Making sense of noninferiority: a clinical and statistical perspective on its application to cardiovascular clinical trials. *Prog Cardiovasc Dis* 2007;49:284–299.

133. Califf RM. A perspective on the regulation of the evaluation of new antithrombotic drugs. *Am J Cardiol* 1998;82(Suppl):25P–35P.
134. SPORTIF Executive Steering Committee for the SPORTIFV Investigators. Ximelagatran vs warfarin for stroke prevention in patients with nonvalvular atrial fibrillation. A randomized trial. *JAMA* 2005;293:690–698.
135. Kaul S, Diamond GA, Weintraub WS. Trials and tribulations of non-inferiority: the ximelegatran experience. *J Am Coll Cardiol* 2005;46:1986–1995.
136. Temple R, Ellenberg SS. Placebo-controlled trials and active-control in the evaluation of new treatments. Part 1: ethical and scientific issues. *Ann Intern Med* 2000;133:455–463.
137. Ellenberg SS, Temple R. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 2: practical issues and specific cases. *Ann Intern Med* 2000;133:464–470.
138. Siegel JP. Equivalence and noninferiority trials. *Am Heart J* 2000;139(Suppl):166–170.
139. Hung JHM, Wang SJ, Tsong Y, et al. Some fundamental issues with non-inferiority testing in active controlled trials. *Stat Med* 2003;22:213–225.
140. Hung HM, Wang SJ, O'Neill R. A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. *Biom J* 2005;47:28–36.
141. D'Agostino RB Sr, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues-the encounters of academic consultants in statistics. *Stat Med* 2003;22:169–186.
142. Blackwelder WC. "Proving the null hypothesis" in clinical trials. *Control Clin Trials* 1982;3:345–353.
143. Hasselblad V, Kong DF. Statistical methods for comparison to placebo in active-control trials. *Drug Info J* 2001;35:435–449.
144. Fleming TR. Current issues in non-inferiority trials. *Stat Med* 2008;27:317–332.
145. Peto R. Why do we need systematic overviews of randomized trials? (Modified transcript of an oral presentation). *Stat Med* 1987;6:233–240.
146. Yusuf S. Obtaining medically meaningful answers from an overview of randomized clinical trials. *Stat Med* 1987;6:281–286.
147. Hennekens CH, Buring JE, Hebert PR. Implications of overviews of randomized trials. *Stat Med* 1987;6:397–402.
148. Simon R. The role of overviews in cancer therapeutics. *Stat Med* 1987;6:389–393.
149. Goodman SN. Meta-analysis and evidence. *Control Clin Trials* 1989;10:188–204.
150. Meinert CL. Meta-analysis: science or religion? *Control Clin Trials* 1989;10(Suppl):257S–263S.
151. Altman L. New method of analyzing health data stirs debate. *New York Times*, August 21, 1990.
152. Sacks HS, Berrier J, Reitman D, et al. Meta-analyses of randomized controlled trials. *N Engl J Med* 1987;316:450–455.
153. DeMets DL. Methods for combining randomized clinical trials: strengths and limitations. *Stat Med* 1987;6:341–348.
154. Cochran WG. Some methods for strengthening the common chi-square tests. *Biometrics* 1954;10:417–451.
155. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959;22:719–748.
156. Thompson SG. Meta-analysis of clinical trials. In *Encyclopedia of Biostatistics*. New York: Wiley, 1998, pp. 2570–2579.
157. Pocock SJ, Hughes MD. Estimation issues in clinical trials and overviews. *Stat Med* 1990;9:657–671.
158. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat Med* 1988;7:889–894.
159. Berlin JA, Laird NM, Sacks HS, Chalmers TC. A comparison of statistical methods for combining event rates from clinical trials. *Stat Med* 1989;8:141–151.
160. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177–188.

161. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Stat Med* 1991;10:1665–1677.
162. Brand R, Kragt H. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Stat Med* 1992;11:2077–2082.
163. Carroll RJ, Stefanski LA. Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. *Stat Med* 1994;13:1265–1282.
164. Higgins JPT, Green S (eds.). *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, UK: John Wiley & Sons, 2008.
165. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA Statement. *PLoS Med* 2009;6:e1000097.
166. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med* 2009;6:e1000100.
167. Chalmers TC, Matta RJ, Smith H, Kunzler AM. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *N Engl J Med* 1977;297:1091–1096
168. May GS, Furberg CD, Eberlein KA, Geraci BJ. Secondary prevention after myocardial infarction: a review of short-term acute phase trials. *Prog Cardiovasc Dis* 1983;25:335–359.
169. Baum ML, Anish DS, Chalmers TC, et al. A survey of clinical trials of antibiotic prophylaxis in colon surgery: evidence against further use of no-treatment controls. *N Engl J Med* 1981;305:795–799.
170. Canner PL. Aspirin in coronary heart disease: comparison of six trials. *Ir J Med Sci* 1983;19:413–423
171. Wang PH, Lau J, Chalmers TC. Meta-analysis of effects of intensive blood-glucose control on late complications of type I diabetes. *Lancet* 1993;341:1306–1309.
172. Himel HN, Liberati A, Gelber RD, Chalmers TC. Adjuvant chemotherapy for breast cancer – A pooled estimate based on published randomized control trials. *JAMA* 1986;256:1148–1159.
173. Yusuf S, Peto R, Lewis J, et al. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Prog Cardiovasc Dis* 1985;27:335–371.
174. Yusuf S, Collins R, Peto R, et al. Intravenous and intracoronary fibronolytic therapy in acute myocardial infarction: overview of results on mortality, reinfarction and side-effects from 33 randomized controlled trials. *Eur Heart J* 1985;6:556–585.
175. Hennekens CH, Buring JE, Sandercock P, et al. Aspirin and other antiplatelet agents in the secondary and primary prevention of cardiovascular disease. *Circulation* 1989; 80:749–756.
176. Goldman L, Feinstein AR. Anticoagulants and myocardial infarction. The problems of pooling, drowning, and floating. *Ann Intern Med* 1979;90:92–94.
177. Furberg CD, Morgan TM. Lessons from overviews of cardiovascular trials. *Stat Med* 1987;6:295–303.
178. Collins R, Gray R, Godwin J, Peto R. Avoidance of large biases and large random errors in the assessment of moderate treatment effects: the need for systematic overviews. *Stat Med* 1987;6:245–250.
179. Wittes RE. Problems in the medical interpretation of overviews. *Stat Med* 1987;6:269–276.
180. Chalmers TC, Levin H, Sacks HS, et al. Meta-analysis of clinical trials as a scientific discipline. I: Control of bias and comparison with large co-operative trials. *Stat Med* 1987;6:315–325.
181. Bailey KR. Inter-study differences: how should they influence the interpretation and analysis of results? *Stat Med* 1987;6:351–358.
182. Furberg CD. Lipid-lowering trials: results and limitations. *Am Heart J* 1994;128:1304–1308.
183. Berlin JA, Begg CB, Louis TA. An assessment of publication bias using a sample of published clinical trials. *J Am Stat Assoc* 1989;84:381–392.
184. Simes RJ. Confronting publication bias: a cohort design for meta-analysis. *Stat Med* 1987;6:11–29.

185. Chalmers TC, Frank CS, Reitman D. Minimizing the three stages of publication bias. *JAMA* 1990;263:1392–1395..
186. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *Br Med J* 1994;309:1351–1355.
187. Johnson RT, Dickersin K. Publication bias against negative results from clinical trials: three of the seven deadly sins. *Nat Clin Pract Neurol* 2007;3:590–591.
188. Chalmers TC. Randomization of the first patient. *Med Clin North Am* 1975;59:1035–1038.
189. Cui L, Hung HM, Wang SJ. Modification of sample size in group sequential clinical trials. *Biometrics* 1999;55:853–857.

Chapter 18
Closeout

The closeout phase starts with the final follow-up visit of the first participant enrolled and lasts until all the analyses have been completed. It is evident that well before the scheduled end of the trial, there needs to be a fairly detailed plan for this phase if the study is to be completed in an orderly manner. Importantly, one must be prepared to implement or modify this plan prior to the scheduled termination since unexpected trial results, either beneficial or harmful, may require the trial to be stopped early.

This chapter addresses a number of topics on the closeout process. Although many of them relate primarily to large single-center or multicenter trials, they also apply to smaller studies. The topics discussed include technical procedures for the termination of the trial, cleanup and verification of data, dissemination of trial results, storage of study material, and poststudy follow-up. Obviously, the details of the closeout plan have to be tailored for each particular trial.

Fundamental Point

The closeout of a clinical trial is usually a fairly complex process that requires careful planning if it is to be accomplished in an orderly fashion.

Termination Procedures

Planning

Many details of a closeout depend on factors that only become known once the trial is underway or participant enrollment has been completed. Nevertheless, general planning for the closeout ought to start early. There are arguments for initiating this process on Day 1 of the trial. One major issue is that the trial may not continue through its scheduled termination. Greater than expected benefit or unexpected

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_18, © Springer Science+Business Media, LLC 2010

harm may lead to early termination. A more subtle reason is that developing plans for closeout after the trial is well underway may be interpreted by the blinded investigators as a signal of imminent trial termination. Thus, another recommendation is to develop the general closeout plans prior to the first meeting of the independent monitoring committee [1].

The closeout phase needs its own written protocol or operating procedures with respect to termination activities, dissemination of results, and data cleanup and storage. Although the literature on the topic of closeout is scant, there are a few good descriptions of the process [2].

Scheduling of Closeout Visits

If each participant in a clinical trial is to be followed for a fixed period of time, the closeout phase will be of the same duration as the enrollment phase. If recruitment took 2 years, the closeout phase would last 2 years. This fixed follow-up design may not be desirable since terminating the follow-up of some participants while others are still being actively followed can create problems. In some blinded trials, the code for each participant is broken during the last scheduled follow-up visit. If the unblinding occurs over a span of many months or years, there is the possibility of the investigator learning information that could suggest the identity of the drugs taken by participants still actively followed in the trial. This may happen even if the drug codes are unique for each participant. The investigator may start associating a certain symptom or constellation of symptoms and signs with particular drug codes.

An alternative and frequently used plan involves following all participants to a shortened closeout period to avoid the problems described above. Another advantage of following this design is the added power of the trial. The follow-up period is extended beyond the minimum time for all but the last participant enrolled. In a trial with 2 years of uniform recruitment, the additional follow-up period would increase by an average of up to 1 year. In addition, this approach might be more cost-efficient when the clinic staff is supported solely by the sponsor of the trial. With all participants followed to a shortened closeout period, full support of personnel can be justified until all the participants have been seen for the last time. In trials where the participants are phased out after a fixed time of follow-up, an increase in the staff/participant ratio may be unavoidable.

Despite the problems with following all participants for a fixed length of time, this approach may be preferable in certain trials, particularly those with a relatively short follow-up phase. In such studies, there may be no realistic alternative. In addition, it may not be logistically feasible to conduct a large number of closeout visits in a short time. Depending on the extent of data collection during the last visit, the availability of staff, and weekly clinic hours, seeing 100–150 participants at a clinic may require a month or two. A decision on the type of follow-up plan should be based on the scientific question as well as logistics.

Final Response Ascertainment

During trial termination, it is important in any trial to obtain, to the extent possible, response variable data on every enrolled participant. It is particularly so in trials where the main response variables are continuous ones such as laboratory data or a performance measure. By necessity, the response variable data must be obtained for each participant at the last follow-up visit because it marks the end of treatment and follow-up. If a participant fails to show up for the last visit, the investigator will have missing data. When the response variable is the occurrence of a specific event, such as a nonfatal stroke, the situation may be different if the information can be obtained without having the participant complete a visit.

If a participant suffers an event after her last follow-up visit, but before all participants have been seen for the final visit, the study must have a firm a priori rule whether or not that response variable should be included in the data analysis. For the participants who complete their participation, the simplest solution is to let the last follow-up visit denote each participant's termination of the trial. For participants who do not show up for the last visit, the study has to decide when to make the final ascertainment. If death is a response variable, vital status is usually determined as of the last day that the participant was eligible to be seen. The counting rule must be clearly specified in the study protocol or in the manual of procedures.

A number of means have been used to track participants and to determine their vital status. These include the use of a person's identification number (e.g., Social Security number in the U.S.) or contact with relatives or employers. In countries with national death registries, including the U.S., mortality surveillance is simpler and probably more complete than in countries without such registries. Agencies that specialize in locating people have been used in several trials. In the Digitalis Investigation Group trial [3], a search agency was used, but the searches were limited to records only. It utilized directory assistance, credit header reports, property records, obituary searches, database mailing lists for magazine subscriptions, and other similar means. No personal contact was allowed. These constraints probably limited the success of finding participants lost to follow-up. This process is very sensitive since a search may be looked upon as an intrusion into the privacy of the participant. The integrity of a trial and the importance of its results plus the participant's initial agreement to participate in the trial have to be weighed against a person's right to protect his or her privacy. Investigators may need to include in the informed consent form a sentence stating that the participant agrees to have his or her vital status determined at the end of the trial even if he or she has by then stopped participating actively or withdrawn his or her general consent. It helps to initiate the process of obtaining information on vital status on inactive participants well in advance of the closeout phase.

The uncertainty of the overall results rises as the number of participants for whom response variable data are missing at trial termination increases. For example, assume that death from any cause is the primary response variable in a trial, and the observed mortality is 15% in one group and 10% in the other group. Depending on study size, this group difference might be statistically significant. However, if 10%

of the participants in each group were lost to follow-up, the observed outcome of the trial may be in question. It cannot be assumed that the mortality experience among those lost to follow-up is the same as for those who stayed in the trial, or that those lost to follow-up in one group have a mortality experience identical to those lost to follow-up in the other group. Equally important, there should be no differential assessment in the study groups. Therefore, every effort should be made to ensure that the final ascertainment of response variables is as complete as possible. Special efforts are required by each clinic to locate participants who withdrew or were lost to follow-up. In the Comparison of Medical Therapy, Pacing, and Defibrillation in Heart Failure (COMPANION) trial [4] of defibrillator versus pacemaker versus best medical care, the withdrawal of consent was 4 times higher in the medical care group than in the other two groups when the trial was terminated and the follow-up ended. At a recommendation by the Data Monitoring Board, the investigators approached the participants who had withdrawn their consent and obtained their permission to collect data on vital status and hospitalizations retrospectively for the duration of the trial. This was done at a substantial extra cost and loss of time.

It is a mistaken concept that when a participant goes off the study medication or intervention, he or she is out of the study and thus no longer followed, or at least not followed beyond some short period of time such as 7 days and 30 days. In the Adenomatous Polyp Prevention on Vioxx (APPROVe) study, the participants who stopped their study medication due to adverse effects and other reasons were not followed beyond 14 days of going off the drug [5]. In the reanalysis, the problem with this "informative censoring" was revealed, and an extra full year of follow-up of all randomized participants after stopping study treatment was added. This analysis showed that the excess number of drug-induced major cardiovascular events observed during active treatment continued to increase during the first year after the treatment was stopped. The adjusted hazard ratio for the extra year was 1.41 (95% CI 0.77–2.59).

Transfer of Posttrial Care

The termination of a long-term study can be difficult due to the bonding that often develops between the participants and the clinic staff. The final visit needs to be carefully planned to deal not only with this issue but also with the obligation to inform the participants of which medication they were on (in a blinded study), their individual study data, and of the overall study findings (often at a later time). Referral of participants to a regular source of medical care is another important issue (see Chap. 2).

If the closeout is extended over a long period, as it would be if each participant were followed for the same duration, any early recommendation to an individual participant would have to be based on incomplete follow-up data which may not reflect the final conclusions of the trial. Moreover, information could "leak out" to

the participants still actively treated, thus affecting the integrity of the trial. Although it is highly desirable to provide each participant with a recommendation regarding continued treatment, doing so may not be possible until the study is completely over, and the trial results are known. When unblinding occurs over a span of months or years, the investigator is in an uncomfortable position of ending a participant's participation in the trial and asking him or her to wait for months before he or she can be informed of the study results and be advised what to do. If the incomplete results are clear-cut, it can be easy to arrive at such recommendations. However, in such an instance, the investigator would be confronted with an ethical dilemma. How can he recommend that a participant start, continue or discontinue a new intervention while keeping other participants active in the trial? For this reason, we prefer a shortened period of trial closeout.

Data and Other Study Material

Cleanup and Verification

Verification of data may be time-consuming, and it can conflict with the desire of the investigator to publish his findings as early as possible. While publication of important information should not be delayed unnecessarily, results should not be put into print before key data have been verified. Despite attempts to collect complete, consistent, and error free data, perfection is unlikely to be achieved. Traditional monitoring systems are likely to reveal missing forms, unanswered items on forms, and conflicting data. In isolated cases, they may also uncover falsification of individual data [6, 7] and, in the worst cases, fabrication of all data on fictitious participants [8–10]. Data cleanup and verification typically continue for months after the completion of closeout visits, though the use of electronic records has reduced the burden of cleanup and verification. It is necessary to be realistic in the cleanup process. This means "freezing" the files at a reasonable time after the termination of participant follow-up and accepting some incomplete data. Obviously, the efforts during cleanup should be directed toward the most critical areas required to answer the primary question and serious adverse events.

We strongly recommend that study forms and data be continuously monitored throughout a trial as pointed out in Chap. 11. Data editing should be initiated as soon as possible because it is difficult to get full staff cooperation after the trial and its funding are over. Early monitoring may reveal systematic problems that can be corrected. Staff feedback is also important. Approaches for Statistical Process Control (SPC) audits are now available, and they have been shown to reduce the overall database error rates significantly [11].

Any clinical trial may have its results reviewed, questioned and even audited. Traditionally, this review has been a scientific one. However, since regulatory and other special interest groups may want to look at the data, the key results should be properly verified, documented, and filed in an easily retrievable manner.

The extent of this additional documentation of important data depends on the design of each trial. Various models have been used for this purpose. A simple model requires each investigator to send a duplicate of all death or major event forms on an ongoing basis to an individual member of the independent monitoring committee. In one multicenter study, the investigators were asked at the end of the follow-up to send a list of all the deceased participants along with the date of death to an office independent of the data coordinating center. In another trial, an outside group of experts audited the data before the results were published. An extreme example employed in one large multicenter trial was the establishment of a second data coordinating center. Duplicates of key study forms were submitted to this center, which generated separate data reports. This approach is obviously costly and, in our view, did not turn out to be worthwhile. Common to all models is an attempt to maintain credibility.

Procedures for data clean-up and verification in trials conducted for regulatory approval add substantially to the trial cost and complexity. Many such trials collect a large quantity of data. Final verification of these data is both time-consuming and costly [12, 13]. As noted in Chap. 11, investigators should, when designing such trials, both limit the amount of data and decide which data are essential and require full final verification.

Storage

Investigators should consider storing various kinds of material after a trial has ended. One set of documents such as trial protocol, manual of procedures, study forms, and analytic material, including electronic records, should be kept by the investigator and sponsor. In addition, a list containing identifying information for all participants who enrolled in a trial ought to be stored at the institution where the investigation took place. Local regulations sometimes require that individual participant data such as copies of study forms, laboratory reports, electrocardiograms, and x-rays be filed for a defined period of time along with the participant's medical records. Storage of these data electronically clearly eases the problem of inadequate space. The actual trial results and their interpretation are usually published and can be retrieved through a library search. Exceptions are obviously findings which never reach the scientific literature. This will hopefully change with the new requirements for reporting key findings of trials registered on http://www.ClinicalTrials.gov. It may also be desirable in these cases to file draft manuscripts along with other documentation and analytic material.

In planning for a new trial, an investigator may want to obtain unpublished data from other investigators who have conducted trials in a similar population or tested the same intervention. Similarly, in preparation of a review article, a meta-analysis or a paper on the natural history of a disease, an investigator may want to obtain additional information from published trials. Tables and figures in printed scientific papers seldom include everything that may be of interest. The situation is changing with

the introduction of online journals, which have no space restrictions. These journals can publish full protocols, forms, manuals, and even raw data [14]. However, no uniform mechanism exists today for getting access to such study material from terminated trials. Even if information is available, it may not be in a reasonable and easily retrievable form. Substantial cooperation is usually required from investigators originally involved in data collection and analysis [15].

The storage of biological material has raised new issues as it relates to genetic analyses. Biospecimens from well-characterized populations followed for long durations in clinical trials are in demand. These can be used to determine whether patient subgroups with a specific genotype are more likely to benefit or to experience serious adverse events. The availability of these specimens for analysis depends on the wording of the informed consent (see Chap. 2). Patient privacy has to be considered as always.

Storage of biomaterials may be costly. Freezers must be maintained, and a system for retrieval of specimens or aliquots without damaging the remaining material must be implemented. Unlike with retrieval and distribution of data, most specimens may only be used once. Therefore, investigators need to develop a system for deciding when and how to use or distribute biospecimens. The cost and benefits, as well as the duration of storage must be considered. Central specimen repositories have been created to which investigators may be able to send their materials.

In summary, most trials collect an excess of study material and it may not make sense to store everything. The investigator has to consider logistics, the length of the storage period and cost. He also has to keep in mind that biological material, for example, deteriorates with time and laboratory methods change.

Dissemination of Results

The reporting of findings from a small single-center trial is usually straightforward. The individual participants are often informed about the results during the last follow-up visit or shortly afterward, and the medical community is informed through scientific publications. However, there are situations that make the dissemination of findings difficult, especially the order in which the various interested parties are informed. Particularly in multicenter studies where the participants are referred by physicians not involved in the trial, the investigators have an obligation to tell these physicians about the conclusions, preferably before they read about them in the newspaper or are informed by their patients. In trials with clinics geographically scattered, investigators may have to be brought together to learn the results. In certain instances, the sponsoring party has a desire to make the findings known publicly at a press conference or through a press release. However, although an early press conference followed by an article in a newspaper may be politically important to the sponsor of the trial, it may offend the participants, the referring physicians, and the medical community. They may all feel that they have the right to be informed before the results are reported in the lay press.

We have had good experiences from the following sequence. First, the study leadership informs the other investigators who, in turn, inform the participants. Second, the private physicians of the participants are also informed, in confidence, of the findings. Third, the results are then published in the scientific literature, after which they may be more widely disseminated in other forums. With most journals now being available electronically, publication can often be timed to coincide with the presentation of results at major scientific meetings.

However, there are sometimes unavoidable long delays between the presentation of trial findings at a scientific meeting and the publication of full trial reports in peer-reviewed journals. The medical community may be placed in a difficult position by having to make treatment decisions if the lay press reports on elements of findings many months prior to the publication of trial data in full. The messages released by the lay press are typically very simple. To minimize this problem three recommendations have been made [16]: (1) "congress organizers should insist that published abstracts contain sufficient data to justify the conclusions of the presentation," (2) "investigators should not present results of any study that is likely to influence clinical management until they are in a position to write a full paper," and (3) "journal editors must be willing... to expedite the publication of such papers." These recommendations are reasonable, but there may be exceptions.

In order to facilitate expedited translation of research results, the National Institutes of Health introduced a data sharing policy in October 2003 [17]. The agency's position is that "Data should be made as widely and freely available as possible while safeguarding the privacy of participants and protecting confidential and proprietary data." The risk of wide dissemination of databases is that other investigators might analyze the available data and arrive at different interpretations of results. However, further analysis and discussion of various interpretations of trial data are usually scientifically sound and ought to be encouraged.

In special situations, when a therapy of public health importance is found to be particularly effective or harmful in a trial sponsored by the National Institutes of Health, physicians and the public need to be alerted in a timely manner. The NIH would promptly post a release on its NIH News website (http://www.nih.gov/news/). When the Adenoma Prevention with Celecoxib trial sponsored by the National Cancer Institute was terminated due to a 2.5-fold increased risk of major fatal and nonfatal cardiovascular events for participants taking celecoxib compared to those on a placebo, the release was issued the day after the decision was made to stop the treatment [18]. Three months later, the results were published in *The New England Journal of Medicine*.

At the NIH, individual institutes may also issue their own press releases. Such a release often coincides with the publication of an article in a medical journal. However, institutes with journal permission have issued brief press announcements prior to journal publication. To avoid criticism from physician groups, an institute may also notify the leadership of relevant medical societies before the release. The United States National Library of Medicine also releases timely scientific news on its Medline Plus website (http://www.nlm.nih.gov/medlineplus/news). These releases are not limited to NIH-sponsored research.

The United States Food and Drug Administration also informs physicians and the public about regulatory actions and news. FDA MedWatch Safety Alerts for Human Medical Products are posted on the website (http://www.fda.gov/medwatch/safety/year). Included are brief summaries of products in question and FDA Alerts. The latter provide recommendations and information for Healthcare Providers as well as information for patients to consider. The agency also issues Public Health Advisories (http://www.fda.gov/cder/drug/advisory), which contains information on particularly serious concerns or risks, for both healthcare providers and consumers, of a drug or a class of drugs.

If a serious adverse event has been uncovered by investigators in a trial, the FDA and other regulatory agencies or the trial sponsor may communicate this information to the medical community and thereby indirectly to the lay public through a Dear Healthcare Provider letter.

Wide dissemination of trial findings to the public by investigators and study sponsors is increasingly common, even if the results are of modest scientific or public health importance. Press releases have become part of highly orchestrated marketing campaigns in both industry and government funded trials. We strongly support making trial results, and indeed data, widely available, with the expectation that broad discussion (and reanalysis as appropriate) will assist clinicians and the public in arriving at appropriate decisions as to the value of a trial's intervention.

As emphasized in Chap. 1, clinical trials ought to be registered. Worldwide, there are a large number of registries [19]. Until the enactment of the FDA Amendments Act (FDAAA) in September, 2007, the registration was limited to design information from the trial protocols [20]. The FDAAA expanded the scope to include a trial results database with information on participant demographics and baseline characteristics, primary and secondary outcomes, and statistical analyses. These data should be posted within 12 months of trial completion. The database should also be linked to publically available information from the FDA website. This would include summary safety and effectiveness data, public health advisories, and action packages for drug approval. Serious and frequent adverse event data observed during a trial are to be added within 2 years.

Poststudy Follow-Up

There are three main reasons for poststudy follow-up. One is to find out how soon treatment-induced changes in laboratory values or symptoms return to pretrial level or status. The effect of the intervention may last long after a drug has been stopped, and abnormalities revealed by laboratory measurements or adverse drug effects may not disappear until weeks after the intervention has ended. Second, for certain drugs, such as beta-blockers and steroids, the intervention should not be stopped abruptly. A tapering of the dosage may require additional clinic visits. Third, clinical events may occur differentially in the study groups after the intervention has been stopped due to lingering drug effects. Drug effects may be seen for weeks or months after

treatment has been stopped or there may be unfavorable withdrawal reactions [5]. These activities are separate from the moral obligation of the investigator to facilitate, when necessary, a participant's return to the usual medical care system, to ensure that study recommendations are communicated to his or her private physician and at times to continue the participant on a beneficial new intervention.

Long-term poststudy follow-up of participants is a rather complex process in most countries. The investigators and the sponsor have to decide on what should be monitored. Mortality surveillance can be cumbersome globally and is worth undertaking only if there is a reasonable expectation of getting an almost complete record of vital status. Usually, the justification for long-term poststudy surveillance is based on a trend or an unexpected finding in the trial or from a finding from another source.

Obtaining information on nonfatal events is even more complicated and, in general, its value is questionable. However, a classical illustration that poststudy follow-up for toxicity can prove valuable is the finding of severe toxic effects attributed to diethylstilbestrol. The purported carcinogenic effect occurred 15–20 years after the drug was administered and occurred in female offspring who were exposed in utero [21]. Similarly, the use of unopposed estrogen has been reported to be associated with an increased risk of endometrial cancer 15 or more years after therapy was stopped [22]. One article reported an association between in utero exposure to valproate, an antiepileptic drug, and impaired cognitive function in offsprings at 3 years of age [23].

In 1978, the results of a trial of clofibrate in people with elevated lipids indicated an excess of cases of cancer in the clofibrate group compared to the control group [24]. The question was raised whether the participants assigned to clofibrate in the Coronary Drug Project also showed an increase in the cancer incidence. This was not the case [25]. Only 3% of the deaths during the trial were cancer-related. Subsequently, the WHO study of clofibrate reported that all cause mortality was increased in the intervention group [26]. At the same time, the Coronary Drug Project investigators decided that poststudy follow-up was scientifically and ethically important, and such a study was undertaken. No increase in cancer incidence was noted in the clofibrate group [27]. A more recent example is the Women's Health Initiative, which extended follow-up for 5 years after it reached its scheduled termination in 2005. The example brings up a question: Should investigators of large-scale clinical trials make arrangements for surveillance in case, at some future time, the need for such a study were to arise? The implementation of any poststudy surveillance plan raises challenges. A key one is to find a way of keeping participants' names and addresses in a central registry without infringement upon the privacy of the individuals. The investigator must also decide, with little evidence, on the optimal duration of surveillance after the termination of a trial (e.g., 2, 5, or 20 years).

Another issue of poststudy surveillance relates to a possible beneficial effect of intervention. In any trial, assumptions must be made with respect to time between initiation of intervention and the occurrence of full beneficial effect. For many drugs, this so-called "lag-time" is assumed to be zero. However, if the intervention is smoking cessation, a lipid lowering drug, or a dietary change and the response variable is coronary mortality, the lag-time might be a year or longer. The problem

with such an intervention is that the maximum practical follow-up may not be long enough for a beneficial effect to appear. Extended surveillance after completion of active treatment may be considered in such studies. At the scheduled termination of the Multiple Risk Factor Intervention Trial, the results favored the Special Intervention group over Usual Care but did not reach statistical significance [28]. Almost 4 years later, a statistically significant effect emerged [29].

The poststudy surveillance in the Coronary Drug Project [27] showed unexpected benefit in one of the intervention groups. At the conclusion of the trial, the participants assigned to nicotinic acid had significantly fewer nonfatal reinfarctions, but no difference in survival was detected. Total mortality, after an average 6.5 years in the trial on drug, plus an additional 9 years after the trial, however, was significantly less in the group assigned to nicotinic acid than in the placebo group. There are several possible interpretations of the CDP finding. It may be that this observation is real and that the benefit of nicotinic acid simply took longer than expected to appear. As one plausible mechanism, the earlier reduction in nonfatal myocardial infarction may have finally affected prognosis. Of course, the results may also be due to chance. A major difficulty in interpreting the data relates to the lack of knowledge about what the participants in the intervention and control groups did with respect to lipid lowering and other regimens in the intervening 9 years. Although there was no reason to expect that there was differential use of any intervention affecting mortality, such could have been the case.

The knowledge of the response variable of interest for almost every participant is required if long-term surveillance after completion of regular follow-up is to be worthwhile. The degree of completeness attainable depends on several factors, such as the response variable itself, the length of surveillance time, the community where the trial was conducted, and the aggressiveness of the investigator. Many of the very large trials have successfully monitored participants (or subsets thereof) after closeout to determine whether behavioral effects of the study intervention have been sustained or participants have adhered to recommendations regarding continued treatment.

References

1. Shepherd R, Macer JL, Grady D. Planning for closeout – from day one. *Contemp Clin Trials* 2008;29:136–139.
2. Pressel SL, Davis BR, Wright Jr. JT, et al. for the ALLHAT Collaborative Research Group. Operational aspects of terminating the doxazosin arm of the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). *Control Clin Trials* 2001;22:29–41.
3. Collins JF, Howell CL, Horney RA, for the Digitalis Investigation Group (DIG) Investigators. Determination of vital status at the end of the DIG trial. *Control Clin Trials* 2003;24:726–730.
4. Bristow MR, Saxon LA, Boehmer J, et al. for the Comparison of Medical Therapy, Pacing, and Defibrillation in Heart Failure (COMPANION) Investigators. Cardiac-resynchronization therapy with or without an implantable defibrillator in advanced chronic heart failure. *N Engl J Med* 2004;350:2140–2150.

5. Baron JA, Sandler RS, Bresalier RS, et al. Cardiovascular events associated with rofecoxib: final analysis of the APPROVe trial. *Lancet* 2008;372:1756–1764.
6. Fisher B, Redmond CK: Fraud in breast-cancer trials. *N Engl J Med* 1994;330:1458–1462.
7. Buyse M, George SL, Evans S, et al. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Stat Med* 1999;18:3435–3451.
8. Sheldon T. Dutch neurologist found guilty of fraud after falsifying 438 case records. *Br Med J* 2002;325:734.
9. Ross DB. The FDA and the case of Ketek. *N Engl J Med* 2007;356:1601–1604.
10. POISE Study Group. Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. *Lancet* 2008;371:1839-1847 (Web attachment 1).
11. Rostami R, Nahm M, Pieper CF. What can we learn from a decade of database audits? The Duke Clinical Research Institute experience, 1997–2006. *Clin Trials* 2009;6:141–150.
12. Eisenstein EL, Lemons II PW, Tardiff BE, et al. Reducing the cost of phase III cardiovascular clinical trials. *Am Heart J* 2005;149:482–488.
13. Eisenstein EL, Collins R, Cracknell BS, et al. Sensible approaches for reducing clinical trial costs. *Clin Trials* 2008;5:75–84.
14. Trials Journal. http://www.trialsjournal.com.
15. Hrynaszkiewicz I, Altman DG. Towards agreement on best practice for publishing raw clinical trial data (Editorial). *Trials* 2009;10:17.
16. Editorial. Reporting clinical trials message and medium. *Lancet* 1994;344:347–348.
17. National Institutes of Health (NIH). Final NIH statement on sharing research data. February 26, 2003. http://grants.nih.gov/grants/policy/datasharing/.
18. US Department of Health and Human Services. NIH News. National Institutes of Health. NIH halts use of COX-2 inhibitor in large cancer prevention trial. December 17, 2004. http://www.nih.gov/news/pr/dec2004/od-17.htm.
19. Zarin DA, Ide NC, Tse T, et al. Issues in the registration of clinical trials. *JAMA* 2007; 297:2112–2120.
20. Zarin DA, Tse T. Moving toward transparency of clinical trials. *Science* 2008;319:1340–1342.
21. Herbst AL, Ulfelder H, Poskanzer DC. Adenocarcinoma of the vagina. Association of maternal stilbestrol therapy with tumor appearance in young women. *N Engl J Med* 1971;284:878–881.
22. Paganini-Hill A, Ross RK, Henderson BE. Endometrial cancer and patterns of use of estrogen replacement therapy: a cohort study. *Br J Cancer* 1989;59:445–447.
23. Meador KJ, Baker GA, Browning N, et al on behalf of the NEAD Study Group. Cognitive function at 3 years of age after fetal exposure to antiepileptic drugs. *N Engl J Med* 2009;360:1597–1605.
24. Report from the Committee of Principal Investigators. A co-operative trial in the primary prevention of ischaemic heart disease using clofibrate. *Br Heart J* 1978;40:1069–1118.
25. The Coronary Drug Project Research Group. Clofibrate and niacin in coronary heart disease. *JAMA* 1975;231:360–381.
26. Committee of Principal Investigators. WHO cooperative trial on primary prevention of ischaemic heart disease using clofibrate to lower serum cholesterol: mortality follow-up. *Lancet* 1980; ii:379–385.
27. Canner PL, Berge KG, Wenger NK, et al. Fifteen year mortality in Coronary Drug Project patients: long-term benefit with niacin. *JACC* 1986;8:1245–1255.
28. Multiple Risk Factor Intervention Trial Research Group. Multiple Risk Factor Intervention Trial. Risk factor changes and mortality results. *JAMA* 1982;248:1465–1477.
29. Multiple Risk Factor Intervention Trial Research Group. Mortality rates after 10.5 years for participants in the Multiple Risk Factor Intervention Trial. Findings related to *a priori* hypotheses of the trial. *JAMA* 1990;263:1795–1801.

Chapter 19
Reporting and Interpreting of Results

The final phase in any experiment is to interpret and report the results. Finding the answer to a challenging question is the goal of any research endeavor. Proper communication of the results to clinicians also provides the basis for advances in medicine [1]. To communicate appropriately, the investigators have to review their results critically and avoid the temptation of overinterpretation. They are in the privileged position of knowing the quality and limitations of the data better than anyone else. Therefore, they have the responsibility for presenting the results clearly and concisely, together with any issues that might bear on their interpretation. Investigators should devote adequate care, time, and attention to this critical part of the conduct of clinical trials. We believe that a policy of "conservative" interpretation and reporting best serves the interests of readers.

A study may be reported in a scientific journal, but publication is in no way an endorsement of its results or conclusions. Even if the journal uses referees to assess each prospective publication, there is no assurance that they have sufficient experience and knowledge of the issues of design, conduct, and analysis to fully judge the reported study [2]. Only the investigators are likely to recognize subtle, or even not so subtle, weaknesses and problems. As pointed out by a former Editor of *The New England Journal of Medicine* [3], "In choosing manuscripts for publication we make every effort to winnow out those that are clearly unsound, but we cannot promise that those we do publish are absolutely true …. Good journals try to facilitate this process [of medical progress] by identifying noteworthy contributions from among the great mass of material that now overloads our scientific communication system. Everyone should understand, however, that this evaluative function is not quite the same thing as endorsement." This point has been illustrated by Ellenberg et al. [4]. The favorable results of a multicenter trial accompanied by a very positive editorial were published in the *New England Journal of Medicine* only 2 weeks before an Advisory Committee of the FDA voted unanimously against recommending that the intervention, a respiratory syncytial virus immune globulin, be licensed. In the end, it is up to the authors to be as objective as possible and the readers of a scientific article to assess it critically and to decide how to make best use of the reported findings.

In this chapter, we discuss guidelines for reporting, interpretation of findings, and publication bias, as well as the answers to three specific questions that should

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_19, © Springer Science+Business Media, LLC 2010

be considered in preparing a report: (a) Did the trial work as planned? (b) How do the findings compare with those from other studies? (c) What are the clinical implications of the findings? A checklist of what should be included in a report of a clinical trial is provided by the Consolidated Standards of Reporting Trials (CONSORT) group [5–8]. Similar guidelines have been prepared for publications of meta-analyses [9].

Fundamental Point

The investigators have an obligation to review their study and its findings critically and to present sufficient information so that readers can properly evaluate the trial.

Any report of a clinical trial should include sufficient methodological information so that the readers can assess the adequacy of the methods employed. The quality of a trial is typically judged based on the thoroughness and completeness of the material and methods sections of the report. Unfortunately, thorough reporting does not always occur. A survey of 253 randomized trials published in five general medicine journals after the revised CONSORT recommendations found that several aspects (e.g., allocation concealment and various components of blinding) were inadequately discussed [10]. Others [11] have noted that eligibility criteria are sometimes poorly described. Wang et al. [12] conducted a survey of subgroup analyses reported in *The New England Journal of Medicine* over a 1-year period. Subgroup analyses were common, but highly variable in completeness of information presented. As a result, *The Journal* implemented guidelines for reporting subgroup analyses [12].

Terms often used in clinical trial reports are misused. Many authors claim that they performed an “intention-to-treat,” or “ITT” analysis, when in fact data from randomized participants have been excluded from the analysis. There may be good reasons why not all data are available, but such an analysis should not be called intention-to-treat. Readers must look carefully despite claims of an ITT analysis. Sometimes, “modified ITT analysis” is used, which is a contradiction. If not all participants and not all follow-up events are accounted for, the report of the analysis should not say “intention-to-treat.” Another term that is misleading is “per protocol analysis.” Authors use that phrase to apply to analyses that omit data from those who fail to adhere fully to the intervention or otherwise leave the study. We consider this to be an unfortunate use of the term, as it implies that such an analysis is the preferred one specified in the protocol. As we have argued in this book, it is almost never the preferred analysis and should not be so specified in the protocol.

Traditional journals impose page limitations, forcing authors to exclude some important information. Online journals that do not have such page limitations are becoming more common. In addition, many print journals allow supplemental material (e.g., details of methods, extra data) to be included in their electronic versions. Therefore, space limitations are no longer justification for withholding pertinent information.

Data sharing among investigators and public access to data and publications have been proposed, and even required by some clinical trials sponsors [13, 14]. The benefits and limitations of these policies are contentious, but all investigators whose trial was funded by an agency requiring data sharing must keep abreast of the requirements.

Guidelines for Reporting

As noted above, guidelines on how to report a clinical trial exist [5–8]. The International Committee of Medical Journal Editors has issued a set of uniform requirements that are endorsed by a large number of journals [15]. One of the guidelines is assurance that the trial has been listed in a formal registry [16]. In addition, journals have their *Instructions for Authors* that address issues on format as well as content.

With the enormous number of scientific articles published annually, it is impossible for clinicians to keep up with the flow of information. Journals to which one subscribes may have online services to help keep identify articles of particular interest. Other online listings of publications in selected areas to which readers can subscribe can also help, but the clinician still has the obligation to review carefully clinical trial publications. More informative abstracts help clinicians who browse through journals on a regular basis. Valid and informative abstracts are important since clinical decisions are often influenced by abstracts alone [17]. For reporting clinical investigations, many journals have adopted the recommendation [18] for structured abstracts, which include information on objective, design, setting, participants, intervention(s), measurements and main results, and conclusion(s). The early experience of structured abstracts was reviewed by Haynes et al., and comments were “supportive and appreciative.” Those authors recommended some modifications of the guidelines [19]. We strongly endorse the now common use of the structured abstract.

Authorship

Decisions of authorship are both sensitive and important [20, 21]. It is critical that decisions are made at an early stage. Cases of scientific fraud have reminded us that being an author carries certain responsibilities and should not be used as a means to show gratitude. Guidelines regarding qualifications for authorship are included in general instructions for manuscripts [15]. *The New England Journal of Medicine* in 1991 instituted guidelines that prohibited group authorship (common to large multicenter studies), restricted authorship to 12 (with a possibility for waiver), and limited the space devoted to acknowledgement [22]. Meinert [23] came to the defense of group authorship and expressed concern over the possible effect of this policy on multicenter work. We believe that group authorship is an important part

of clinical trials research. Fairness and equity require proper crediting to those who have made major contributions to the design, conduct, and analysis, not just the few that served on the writing group. A compromise accepted by many journals and recommended by the International Committee of Medical Journal Editors is to allow group authorship but list those who served on the writing committee. As stated in the document from that group [15], some journals ask about the contributions of each person listed as an author or member of a writing group.

Ghost authorship, or the failure to properly credit those who wrote or coauthored a manuscript or who otherwise played a major role in the trial such that they deserve notice, has received considerable attention. Gøtzsche and colleagues [24] conducted a survey of 44 industry-initiated trials and found evidence of ghost authorship in three quarters of the publications. Ross et al. [25] describe publications concerning rofecoxib that were written by the industry sponsor's employees, who were not acknowledged as authors.

The flip side of ghost authorship is guest authorship, where usually highly respected investigators who had little or no role in the writing of the manuscript are given visible authorship. We deplore both of these practices.

Disclosure of Conflict of Interest

Many journals have policies requiring clear statements of possible conflicts of interest [26]. The "Uniform Requirements for Manuscripts Submitted to Biomedical Journals" [15] contains guidelines regarding disclosure of potential conflicts related to individual authors and to the role of the sponsor of the trial. Authors must be forthcoming in disclosing any potential conflicts, as they can affect how readers interpret study findings. Unfortunately, there have been instances where important conflicts were not disclosed and were subsequently discovered [27, 28]. These cases serve both to embarrass the investigators and perhaps unfairly tarnish good research; a situation that could have been avoided had openness been followed in the beginning. We recommend that all authors disclose freely all real, potential, or apparent conflicts of interest.

Presentation of Data

Presentation of the data analysis is important [29–37]. There is a common misunderstanding of the meaning of *p*-values. Only about one-fifth of the respondents to a multiple choice question understood the proper meaning of a *p*-value [38]. The *p*-value tells us how likely an observed difference may have occurred by chance. It conveys information about the level of doubt, not the magnitude of clinical importance of this difference. A *p*-value of 0.05 in a very large trial may be weak evidence of an effect while in a small sample it can be quite strong evidence [30].

The point estimate (the observed result) with its 95% confidence interval (CI) provides us with the best estimates of the size of a difference. The width of the CI is another measure of uncertainty. The *p*-value and the CI are inherently related; thus, if the 95% CI of the difference excludes 0, the difference is statistically significant with $p < 0.05$. The CI permits the readers to use their own value for the smallest clinically important difference in making treatment decisions [29]. Some journals have taken the lead and now require more extensive use of CIs. We advocate reporting of *p*-values, point estimates, and CIs for the major results. They all convey important information and help in evaluating a trial's result.

Interpretation

Many articles have been written to help clinicians in their appraisal of a clinical study [39–44]. Readers should be aware that many publications have deficiencies and can even be misleading. Pocock [45] has given three reasons why readers need to be cautious: (a) some authors produce inadequate trial reports, (b) journal editors and referees allow them to be published, (c) journals favor positive findings. For example, a review of trials of antibiotic prophylaxis found that 20% of the abstracts omitted important information or implied unjustified conclusions [46]. Pocock and coworkers [47] examined 45 trials and concluded that the reporting "appears to be biased toward an exaggeration of treatment differences" and that there was an overuse of significance levels. In a 1982 report, statistical errors were uncovered in a large proportion of 86 controlled trials in obstetrics and pediatrics journals and only 10% of the conclusions were considered justified [48]. In 76% of 196 trials of nonsteroidal anti-inflammatory drugs in rheumatoid arthritis, "doubtful or invalid statements" were found [49]. As mentioned in Chap. 9, inadequate reporting of the methods of randomization and baseline comparability was found in 30–40% of 80 randomized clinical trials in leading medical journals [50]. The criteria for tumor response from articles published in three major journals were incompletely reported, variable, and contributed to the wide variations in reported response rates [51].

Baar and Tannock [52] constructed a hypothetical trial and reported its results in two separate articles: one with errors of reporting and omissions similar to those "extracted from" leading cancer journals and the other with appropriate methods. This exercise illustrates how the same results can be interpreted and reported differently.

The way in which results are presented can affect treatment decisions [53–55]. Almost half of a group of surveyed physicians were more impressed and indicated a higher likelihood of treating their patients when the results of a trial were presented as a relative change in outcome rate compared to an absolute change (difference in the incidence of the outcome event) [54]. A relative treatment effect is difficult to interpret without knowledge of the event rate in the comparison group. The use of a "summary measure," such as the number of persons who need to be treated to prevent one event, had the weakest impact on clinicians' views of therapeutic effectiveness [55]. We recommend that authors report both absolute and relative changes in outcome rates.

Publication Bias

Timely preparation and submission of the trial results – whether positive, neutral, or negative – ought to be every investigator's obligation. The written report is the public forum that all the work of a clinical trial finally faces. Regrettably, negative trials are more likely to remain unpublished than positive trials. The first evidence of this publication bias came from a survey of the psychological literature. Sterling [56] noted in 1959 that 97% of 294 articles involving hypothesis testing reported a statistically significant result. The situation was similar for medical journals decades later; about 85% of articles – clinical trials and observational studies – reported statistically significant results [57]. Simes [58] compared the results of published trials with those from trials from an international cancer registry. A pooled analysis of published therapeutic trials in advanced ovarian cancer demonstrated a significant advantage for a combination therapy. However, the survival ratio was lower and statistically nonsignificant when the pooled analysis was based on the findings of all registered trials. Several surveys have identified selective reporting and/or multiple publications of the same trial [59–63]. Even multicenter trials conducted at a major academic center remained unpublished over 40% of the time. Those trials sponsored by government were published only modestly more often than those sponsored by industry [62].

Turner et al. [63] looked at 74 studies of antidepressant agents that had been registered with the U.S. Food and Drug Administration. Twenty three of the trials had not been published. In addition, those that were published claimed to show results more positive toward the intervention than did a subsequent FDA analysis of the data. Perlis et al. found that financial conflict of interest was common in clinical trials in psychiatry and was associated with clinical trial results that were highly favorable to the intervention [64]. According to Chan and colleagues [61], there were frequent discrepancies between the primary response variable mentioned in the trial protocol and that reported in the publication of results. It has been shown that many abstracts are never followed by full publications [65]. Dickersin et al. [66] found that among 178 unpublished trials with a trend specified, 14% favored the new therapy compared to 55% among 767 published reports ($p < 0.001$). Analysis of factors associated with this bias are, in addition to neutral and negative findings, small sample size and possibly pharmaceutical source of funding [66]. Rejection of a manuscript by a journal is an infrequent reason [67, 68]. However, authors are no doubt aware that it is difficult to publish neutral results and manuscripts may never get written only to be rejected. A survey of the reference lists of trials of nonsteroidal anti-inflammatory drugs revealed a bias toward references with positive outcomes [69].

Selective reporting is viewed as a serious issue. In a survey of clinical trialists, selective reporting was considered among the two most important forms of scientific misconduct [70]. Investigators have the primary responsibility for ensuring that they do not engage in this practice. Journals too have a responsibility to encourage full and honest reporting. They ought to select trials for publication according to the quality

of their conduct rather than according to whether the *p*-value is significant. We expect that the common use of clinical trial registries will encourage more complete reporting of trial results, as those trials begun but not reported are more easily identified.

Did the Trial Work as Planned?

Baseline Comparability

The foundation of any clinical trial is the effort to make sure that the study groups are initially comparable so that differences between the groups over time can be reasonably attributed to the effect of the intervention. Randomization is the preferred method used to obtain baseline comparability. The use of randomization does not necessarily guarantee balance at baseline in the distribution of known or unknown prognostic factors. Baseline imbalance is fairly common in small trials but may also exist in large trials (see Chap. 9). Therefore, both a detailed description of the randomization process, including efforts made to prevent prior knowledge on the part of the investigator of the intervention assignment, and an evaluation of baseline comparability are essential. Should the trial be nonrandomized, the credibility of the findings hinges even more upon an adequate documentation of this comparability. For each group, baseline data should include means and standard deviations of known and possible prognostic factors. Note that the absence of a statistically significant difference for any of these factors does not mean that the groups are balanced. In small studies, large differences are required in order to reach statistical significance. In addition, small trends for individual factors can have an impact if they are in the same direction. A multivariate analysis to evaluate balance may be advantageous. Of course, the fact that major prognostic factors may be unknown will produce some uncertainty with regard to baseline balance. Adjustment of the findings on the basis of observed baseline imbalance should be performed and any difference between unadjusted and adjusted analyses should be carefully explained.

Blindness

Double-blindness is a desirable feature of a clinical trial design because, as already discussed, it diminishes bias in the assessment of response variables that require some element of judgment. However, many studies are not truly double-blinded to all parties from start to finish. While an individual side-effect may be insufficient to unblind the investigator, a constellation of effects often reveals the group assignment. A specific drug effect such as a marked fall in blood pressure in an antihypertensive drug trial – or the absence of such an effect – might also indicate which is the treatment group. Although the success of blinding may be difficult for the

investigator to assess, an evaluation should be done. Readers of a publication ought to be informed about the degree of unblinding. An evaluation such as the one provided by Karlowski and colleagues for a trial of vitamin C [71] is commendable. In a completed double-blind, placebo-controlled trial of a lipid-lowering agent, the participants were asked at the close-out visit whether they had their lipids analyzed during the 3-year treatment period. Although unblinding was discouraged, over half of them admitted that they had done so. It is possible that information on the lipid values could have led to an increased cross-over rate in the placebo group.

Adherence and Concomitant Treatment

In estimating sample size, investigators often make assumptions regarding the rate of nonadherence. Throughout follow-up, efforts are made to maintain optimal adherence to the intervention under study and to monitor adherence. When interpreting the findings, one can then gauge whether the initial assumptions were borne out by what actually happened. When adherence assumptions have been too optimistic, the ability of the trial to test adequately the primary question may be less than planned. The study results must be reported and discussed with the power of the trial in mind. In trials showing a beneficial effect of a specific intervention, nonadherence is usually a minor concern. Two interpretations of the effect of nonadherence are possible. It may be argued that the intervention would have been even more beneficial had adherence been higher. On the other hand, if all participants (including those who for various reasons did not adhere entirely to the dosage schedule of a trial) had been on full dose, there could have been further adverse events or toxic effects in the intervention group.

Also of interest is the comparability of groups during the follow-up period with respect to concomitant interventions. The use of drugs other than the study intervention, changes in lifestyle, and general medical care – if they affect the response variable – need to be measured. Of course, as mentioned in Chap. 17, adjustment on postrandomization variables is inappropriate. As a consequence, when imbalances exist, the study results must be interpreted cautiously.

Limitations

When the results of a "superiority" trial (i.e., one in which an intervention is evaluated to see if it differs from a control) indicate no statistically significant difference between the study groups, there are several possible explanations. The dose of the studied intervention may have been too low or too high; the technical skills of those providing the intervention (e.g., surgical procedure) may have been inadequate; the sample size may have been too small, giving the trial insufficient power to test the hypothesis (Chap. 8); there may have been major adherence problems; concomitant interventions may have

reduced the effect that would otherwise have been seen; or the outcome measurements may not have been sensitive enough or the analyses may have been inadequate. Finally, chance is another obvious explanation. The authors should provide the readers with enough information in the methods and results sections for them to judge for themselves why an intervention may not have worked. In the discussion section, the authors should also offer their best understanding of why no difference was found.

For equivalence or noninferiority trials, inadequate design or conduct, or poor adherence on the part of participants, can lead to what the investigators and sponsors consider as the "desired" outcome, that is, no discernable difference between intervention groups. Perhaps even more than in superiority trials, the authors must recognize and clearly acknowledge any study limitations that could have contributed to the lack of difference. In some cases, an "on treatment" analysis might be warranted, in addition to the intention-to-treat analysis.

What are limitations of the findings? One needs to know the degree of completeness of data in order to evaluate a trial. A typical shortcoming, particularly in long-term trials, is that the investigator may lose track of some participants or for other reasons have missing data. These participants are usually different from those who remain in the trial, and their event rate or outcome measurements may not be the same. Vigorous attempts should be made to keep the number of persons lost to follow-up to a minimum. The credibility of the findings may be questioned in trials in which the number of participants lost to follow-up is large in relation to the number of events. A conservative approach in this context is to assume the "worst case." This approach assumes the occurrence of an event in each participant lost to follow-up in the group with lower incidence of the response variable, and it assumes no events in the comparison group. After application of the "worst case" approach, if the overall conclusions of the trial remain unchanged, they are strengthened. However, if the worst-case analysis changes the conclusions, the trial may have less credibility. Other approaches to handling missing outcome data are discussed in Chap. 17. The degree of confidence in the conclusion will depend upon the extent to which the outcome could be altered by the missing information.

Analysis

As addressed in Chap. 17, results may be questionable if participants randomized into a trial are withdrawn from the analysis. Withdrawal after randomization undermines the goal of conducting a valid, unbiased trial. It should be avoided. Investigators who support the concept of allowing withdrawals from the analysis should be required to report analyses both with, and without, withdrawals. If both analyses give approximately the same result, the findings are confirmed. However, if the results of the two analyses differ, believe the intention-to-treat analysis while exploring the reasons for the differences.

In evaluating possible benefit of an intervention, more than one response variable is often assessed which raises the issue of multiple comparisons (Chap. 17). In essence, the

chance of finding a nominally statistically significant result increases with the number of comparisons. This is true whether there are multiple response variables, repeated comparisons for the same response variable, subgroup analyses or whether various combinations of response variables are tested. In the survey of 45 trials in three leading medical journals, the median number of significance tests per trial was eight; more than 20 tests were reported in six trials [46]. The potential impact of this multiple testing on the findings and conclusion of a trial ought to be considered. A conservative approach in the interpretation of statistical tests is again recommended. When several comparisons have been made, a more extreme statistic might be required before a statistically significant difference could be claimed. One approach is to require a p-value <0.01 for a limited number of secondary outcomes in order to declare a treatment difference statistically significant. An alternative approach is to consider all of the subsidiary analyses exploratory and hypothesis generating [46]. Authors of a report should indicate the total number of comparisons made during a trial and in the analysis phase (not just those selected for reporting). Readers should focus attention on p-values for protocol-specified comparisons.

The main objective of any trial is to answer the primary question. Findings related to one of the secondary questions may be interesting, but they should be put in the proper perspective. Are the findings for the related primary and secondary response variables consistent? If not, attempts ought to be made to explain discrepancies. Explaining inconsistencies was particularly important in the Cooperative Trial in the Primary Prevention of Ischaemic Heart Disease [72]. In that trial, the intervention group showed a statistically significant reduction in the incidence of major ischemic heart disease (primary response variable) but a significant increase in mortality for any cause (secondary response variable).

In all studies, evidence for possible serious adverse events from the intervention needs to be presented. Comparison of adverse events among those participants who adhered to the intervention may provide a more conservative assessment, in the sense that it leans toward safety. Authors might consider analyzing adverse event data both using intention-to-treat and on-treatment approaches. In the final conclusion, the overall benefit should be weighed against the risk of harm. This assessment of the balance, however, is too infrequently done (Chap. 12).

How Do the Findings Compare with Results from Other Studies?

The findings from a clinical trial should be placed in the context of current knowledge. Are they consistent with knowledge of basic science, including presumed mechanism of action of the intervention? Although the precise mechanism may be unclear, when the outcome can be explained in terms of known biological actions, the conclusions are strengthened. Do the findings confirm the results of studies with similar interventions or different interventions in similar populations?

It is important here to keep in mind that a substantial proportion of initiated and even completed trials are never published. Additionally, a review of the completeness

of articles cited in reference lists of clinical trial publications suggests that studies with neutral or negative results tend not to be cited [68]. Among published trials the response to a given drug or drug combination can vary markedly [52, 73, 74]. The reported response rates to fluorouracil therapy varies between 8 and 85% for metastatic colorectal cancer [52]. Much of this variation may be explained by differences in participant selection, including genetic variation, treatment regimen, and concomitant intervention, but major differences may also reflect the way the data were analyzed and reported. In a review of 51 randomized clinical trials in congestive heart failure, the authors attributed conflicting results to lack of uniform diagnostic criteria [73]. In a thoughtful editorial, Packer [74] pointed out that several other factors could explain discordant results. He suggested that the characteristics of the enrolled participants may be more important than the definition of congestive heart failure. Differences in design – sample size, dose, and duration of intervention – may affect the trial findings. Other factors might be differences in criteria of efficacy and publication policy. Results of positive trials tend to be published several times, for example, both in a regular journal report and in a journal supplement funded by the pharmaceutical industry. Bero et al. [75] analyzed the symposium issues of 11 journals and concluded that the number increased steadily between 1966 and 1989, that they often had promotional attributes, were less likely to be peer-reviewed, and were more likely to have misleading titles.

Generally, credibility of a particular finding increases with the proportion of good independent studies that come to the same conclusion. Inconsistent results are not uncommon in research. In such cases, the problem for both the investigators and the readers is to try to determine the true effect of an intervention. How and why results differ need to be explored. The use of confidence limits has the advantage of allowing the readers to compare findings and assess whether the results of different trials could, in fact, be consistent.

What are the Clinical Implications of the Findings?

It is appropriate, of course, to generalize the results to the study population, that is, those people who would have been eligible for and could have participated in the trial. The next step, suggesting that the trial results be applied to a more general population (the majority of which would not even meet the eligibility criteria of the trial) is more tenuous. Readers must judge for themselves whether or not such an extrapolation is appropriate. As seen in Fig. 4.1 in Chap. 4, there is often a considerable winnowing from the initial study population to the final sample. A similar argument applies to the intervention itself. How general are the findings? If the intervention involved a special procedure, such as surgery or counseling, is its application outside the trial setting likely to produce the same response? In a drug trial, the question of dose–effect relationship is often raised. Would a higher dose of the drug have given different results? Can the same claims be made for different drugs that have a similar structure or pharmacological action? Can the results of an

intervention be generalized even more broadly? For example, there have been many trials comparing different statins in the prevention of coronary disease sequelae. If the goal LDL-cholesterol is the same in the groups being compared, should one expect similar outcomes? Based on the experience with cerivastatin [76], statins are unlikely to be the same, at least with respect to adverse events. One problem in trials of devices is that the devices are constantly being modified or improved, with respect to the technology or the software algorithm. Does the trial using the old model have any implications for the latest model or the model to come in the future? For a further discussion of generalization, see Chap. 4.

In 1987, a review found that the majority of therapeutic interventions had not been properly tested in randomized clinical trials [77]; approval may have been granted on the basis of surrogate endpoints or drugs may have multiple indications, only some of which are proven. As discussed in this book, there continue to be examples of drugs that had been approved but when assessed in an adequately designed clinical trial turn out not to be as wonderful as hoped. Skillful marketing has a major impact on practice patterns. The marked regional differences in drug sales cannot be explained on the basis of science, since regions, in principle, have access to the same scientific information. It is difficult to tease out the impact of clinical trials on medical practice from other factors such as marketing and treatment guidelines. There are several examples of trials that have changed practice patterns [78, 79]. Similarly, there are examples where practice was predominantly influenced by the other factors [80].

As with all research, a clinical trial will often raise as many questions as it answers. Suggestions for further research should be discussed. Finally, the investigator might allude to the social, economic, and medical impact of the study findings. How many lives can be saved? How many working days will be gained? Can symptoms be alleviated? Economic implications or cost-effectiveness are important. Any benefit has to be weighed against the cost and feasibility of use in routine medical practice rather than in the special setting of a clinical trial.

References

1. Comroe JH Jr. The road from research to new diagnosis and therapy. *Science* 1978;200:931–937.
2. Glantz SA. Biostatistics: how to detect, correct and prevent errors in the medical literature. *Circulation* 1980;61:1–7.
3. Relman AS. What a good medical journal does. *New York Times*, March 19, 1978; Section IV; p. 22.
4. Ellenberg SS, Epstein JS, Fratantoni JC, et al. A trial of RSV immune globulin in infants and young children: the FDA view. *N Engl J Med* 1994;331:203–204.
5. Moher D, Schulz KF, Altman DG, for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001;357:1191–1194.
6. Altman DG, Schulz KF, Moher D, et al, for the CONSORT Group. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;134:663–694.

7. Ioannidis JPA, Evans SJW, Gøtzsche PC, et al, for the CONSORT Group. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med* 2004;141:781–788.
8. Piaggio G, Elbourne DR, Altman DG, et al, for the CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. *JAMA* 2006;295:1152–1160.
9. Moher D, Liberati A, Tetziaff J, Altman DG, The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009;6:e1000097.
10. Mills EJ, Wu P, Gagnier J, Devereaux PJ. The quality of randomized trial reporting in leading medical journals since the revised CONSORT statement. *Contemp Clin Trials* 2005;26:480–487.
11. Van Spall HGC, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. *JAMA* 2007;297:1233–1240.
12. Wang R, Lagakos SW, Ware JH, et al. Special Report: Statistics in medicine – reporting of subgroup analyses in clinical trials. *N Engl J Med* 2007;357:2189–2194.
13. National Institutes of Health Public Access. http://publicaccess.nih.gov/
14. Zarin DA, Tse T. Moving toward transparency of clinical trials. *Science* 2008;319:1340–1342.
15. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals: writing and editing for biomedical publication (updated October 2008). http://www.icmje.org/.
16. International Committee of Medical Journal Editors. Is this clinical trial fully registered?: a statement from the International Committee of Medical Journal Editors. http://www.icmje.org/clin_trialup.htm
17. Haynes RB, McKibbon KA, Walker CJ, et al. Online access to MEDLINE in clinical settings: a study on the use and usefulness. *Ann Intern Med* 1990;112:78–84.
18. Ad hoc Working Group for Critical Appraisal of the Medical Literature. A proposal for more informative abstracts of clinical articles. *Ann Intern Med* 1987;106:598–604.
19. Haynes RB, Mulrow CD, Huth EJ, et al. More informative abstracts revisited. *Ann Intern Med* 1990;113:69–76.
20. Huth EJ. Preparing to write. In: *How to write and publish papers in medical sciences.* Philadelphia: ISI Press, 1982, pp. 37–40.
21. Huth EJ. Guidelines on authorship of medical papers. *Ann Intern Med* 1986;104:269–274.
22. Kassirer JP, Angell M. On authorship and acknowledgments. *N Engl J Med* 1991;325:1510–1512.
23. Meinert CL. In defense of the corporate author for multicenter trials. *Control Clin Trials* 1993;14:255–260.
24. Gøtzsche PC, Hróbjartsson A, Johansen HK, et al. Ghost authorship in industry-initiated randomised trials. *PLoS Med* 2007; 4:e19.
25. Ross JS, Hill KP, Egilman DS, Krumholz HM. Guest authorship and ghostwriting in publications related to rofecoxib: a case study of industry documents from rofecoxib litigation. *JAMA* 2008;299:1800–1812.
26. Drazen JM, Van Der Weyden MB, Sahni P, et al. Editorial: Uniform format for disclosure of competing interests in ICMJE journals. *N Engl J Med* 2009;361:1896–1897 http://content.nejm.org/cgi/reprint/NEJMe0909052.pdf?resourcetype=HWCIT
27. DeAngelis CD, Fontanarosa PB. Editorial: Resolving unreported conflicts of interest. *JAMA* 2009;302:198–199.
28. Weinfurt KP, Seils DM, Tzeng JP, et al. Consistency of financial interest disclosures in the biomedical literature: the case of coronary stents. *PLoS One* 2008; 3:e2128.
29. Berry G. Statistical significance and confidence intervals. *Med J Aust* 1986;144:618–619.
30. Gardner MJ, Altman DG. Confidence intervals rather than *p*-values: estimation rather than hypothesis testing. *Br Med J* 1986;292:746–750.
31. Simon R. Confidence intervals for reporting results of clinical trials. *Ann Intern Med* 1986;105:429–435.
32. Bulpitt CJ. Confidence intervals. *Lancet* 1987;i:494–497.

33. Braitman LE. Confidence intervals extract clinically useful information from data. *Ann Intern Med* 1988;108:296–298.
34. Freeman PR. The role of p-values in analysing trial results. *Stat Med* 1993;12:1443–1452.
35. Braitman LE. Statistical estimates and clinical trials. *J Biopharm Stat* 1993;3:249–256.
36. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med* 1994;121:200–206.
37. Pocock SJ, Ware JH. Translating statistical findings into plain English. *Lancet* 2009;373:1926–1928.
38. Wulff HR, Andersen B, Brandenhoff P, Guttler F. What do doctors know about statistics? *Stat Med* 1987;6:3–10.
39. Haynes RB, McKibbon KA, Fitzgerald D, et al. How to keep up with the medical literature. *Ann Intern Med* 1986;105:149–153, 309–312, 474–478, 636–640, 810–816, 978–984.
40. Moon TE. Interpretation of cancer prevention trials. *Prev Med* 1989;18:721–731.
41. Fowkes FGR, Fulton PM. Critical appraisal of published research: introductory guidelines. *Br Med J* 1991;302:1136–1140.
42. Cuddy PG, Elenbaas RM, Elenbaas JK. Evaluating the medical literature. *Ann Emerg Med* 1998;12:549–555, 610–620, 679–686.
43. Oxman AD, Sackett DL, Guyatt GH, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature: I. How to get started. *JAMA* 1993;270:2093–2095.
44. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature: II. How to use an article about therapy or prevention: A. Are the results of the study valid? *JAMA* 1993;270:2598–2601.
45. Pocock SJ. *Clinical trials. A practical approach*. Chichester: John Wiley & Sons, 1983.
46. Evans M, Pollock AV. Trials on trial. A review of trials of antibiotic prophylaxis. *Arch Surg* 1984;119:109–113.
47. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. A survey of three medical journals. *N Engl J Med* 1987;317:426–432.
48. Altman DG. Statistics in medical journals. *Stat Med* 1982;1:59–71.
49. Gøtzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal anti-inflammatory drugs in rheumatoid arthritis. *Control Clin Trials* 1989;10:31–56.
50. Altman DG, Doré CJ. Randomization and baseline comparisons in clinical trials. *Lancet* 1990;335:149–153.
51. Tonkin K, Tritchler D, Tannock I. Criteria of tumor response used in clinical trials of chemotherapy. *J Clin Oncol* 1985;3:870–875.
52. Baar J, Tannock I. Analyzing the same data in two ways: a demonstration model to illustrate the reporting and misreporting of clinical trials. *J Clin Oncol* 1989;7:969–978.
53. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med* 1988;318:1728–1733.
54. Forrow L, Taylor WC, Arnold RM. Absolutely relative: how research results are summarized can affect treatment decisions. *Am J Med* 1992;92:121–124.
55. Naylor CD, Chen E, Strauss B. Measured enthusiasm: does the method of reporting trial results alter perceptions of therapeutic effectiveness? *Ann Intern Med* 1992;117:916–921.
56. Sterling TD. Publication decisions and their possible effects on inferences drawn from test of significance – or vice versa. *J Am Stat Assoc* 1959;54:30–34.
57. Dickersin K, Min YI. Publication bias: the problem that won't go away. *Ann NY Acad Sci* 1993;703:135–146.
58. Simes RJ. Publication bias: the case for an international registry of clinical trials. *J Clin Oncol* 1987;4:1529–1541.
59. Melander H, Ahlqvist-Rastad J, Meijer G, Beermann B. Evidence b(i)ased medicine – selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. *Br Med J* 2003;326:1171–1173.

60. Chan A-W, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *Br Med J* 2005;330:753. Epub 2005 Jan 28.
61. Chan A-W, Hróbjartsson A, Haahr MT, et al. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004;291:2457–2465.
62. Turer AT, Mahaffey KW, Compton KL, et al. Publication or presentation of results from multicenter clinical trials: evidence from an academic medical center. *Am Heart J* 2007;153:674–680.
63. Turner EH, Matthews AM, Linardatos E, et al. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med* 2008;358:252–260.
64. Perlis RH, Perlis CS, Wu Y, et al. Industry sponsorship and financial conflict of interest in the reporting of clinical trials in psychiatry. *Am J Psychiatry* 2005;162:1957–1960.
65. Goldman L, Loscalzo A. Fate of cardiology research originally published in abstract form. *N Engl J Med* 1980;303:255–259.
66. Dickerson K, Chan S, Chalmers TC, et al. Publication bias and clinical trials. *Control Clin Trials* 1987;8:343–353.
67. Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA* 1990;263:1385–1389.
68. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991;337:867–872.
69. Gøtzsche PC. Reference bias in reports of drug trials. *Br Med J* 1987;295:654–656.
70. Al-Marzouki S, Roberts I, Marshall T, Evans S. The effect of scientific misconduct on the results of clinical trials: a Delphi survey. *Contemp Clin Trials* 2005;26:331–337.
71. Karlowski TR, Chalmers TC, Frenkel LD, et al. Ascorbic acid for the common cold: a prophylactic and therapeutic trial. *JAMA* 1975;231:1038–1042.
72. Report from the Committee of Principal Investigators. Cooperative trial in the primary prevention of ischaemic heart disease using clofibrate. *Br Heart J* 1978;40:1069–1118.
73. Marantz PR, Alderman MH, Tobin JN. Diagnostic heterogeneity in clinical trials for congestive heart failure. *Ann Intern Med* 1988;109:55–61.
74. Packer M. Clinical trials in congestive heart failure: why do studies report conflicting results. *Ann Intern Med* 1988;109:3–5.
75. Bero LA, Galbraith A, Rennie D. The publication of sponsored symposiums in medical journals. *N Engl J Med* 1992;327:1135–1140.
76. Psaty BM, Furberg CD, Ray WA, Weiss NS. Potential for conflict of interest in the evaluation of suspected adverse drug reactions: use of cerivastatin and risk of rhabdomyolysis. *JAMA* 2004;292:2622–2631.
77. Fineberg HV. Clinical evaluation: how does it influence medical practice? *Bull Cancer* 1987;74:333–346.
78. Collins R, Julian D. British Heart Foundation surveys (1987 and 1989) of United Kingdom treatment policies for acute myocardial infarction. *Br Heart J* 1991;66:250–255.
79. Lamas GA, Pfeffer MA, Hamm P, et al., The SAVE Investigators. Do the results of randomized clinical trials of cardiovascular drugs influence medical practice? *N Engl J Med* 1992;327:241–247.
80. Manolio TA, Cutler JA, Furberg CD, et al. Trends in pharmacologic management of hypertension in the United States. *Arch Intern Med* 1995;155:829–837.

Chapter 20
Multicenter Trials

A multicenter trial is a collaborative effort that involves more than one independent center in the tasks of enrolling and following study participants. Early contributions to the design of these trials were made by Hill [1], and a general discussion of methods was provided by Greenberg [2].

There has been a dramatic increase in the number of multicenter, indeed, multinational, trials in the last three decades. Of course, the sizes of these have varied, depending on the requirements of the study. Multicenter studies are more difficult and more expensive to perform than single-center studies, and they bring perhaps less professional reward due to the need to share credit among many investigators. Nevertheless, they are carried out because single sites cannot enroll enough participants [3]. Over 35 years ago, Levin and colleagues provided many examples of "the importance and the need for well-designed cooperative efforts to achieve clinical investigations of the highest quality" [4].

The reasons for conducting multicenter trials apply even more today, with much of medicine being global in scope. It is common for large late-phase trials sponsored by industry to include a wide geographical representation. Several hundred sites might be involved, each site entering anywhere from several to a few dozen participants. While such dispersion of sites presents logistical challenges for training of personnel and data quality control, the benefits of rapid participant recruitment have generally outweighed these challenges.

Much of the ground work for the development, organization, and conduct of a multicenter trial, was laid in the Coronary Drug Project [5]. A detailed description of multicenter trials is given by Meinert [6]. This chapter discusses the reasons why such studies are conducted and briefly reviews some steps in their planning, design, and conduct.

Fundamental Point

Anyone responsible for organizing and conducting a multicenter study should have a full understanding of the complexity of the undertaking. Problems in conduct of the trial most often originate from inadequate and unclear communication between the participating investigators, all of whom must agree to follow a common study protocol.

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-1-4419-1586-3_20, © Springer Science+Business Media, LLC 2010

Reasons for Multicenter Trials

1. The main rationale for multicenter trials is to recruit adequate number of participants within a reasonable time. Many clinical trials have been – and still are – performed without a good estimate of the number of participants likely to be required to test adequately the main hypothesis. Yet, if the primary response variable is an event that occurs relatively infrequently, or small group differences are to be detected, sample size requirements will be large (Chap. 8).

   Studies requiring hundreds of participants usually cannot be done at one center. In a now old, but still instructive example, the Aspirin Myocardial Infarction Study [7] used 30 centers to enroll the necessary 4,200 participants with a history of a heart attack in 1 year and follow them for an additional 3 years. The largest of these centers enrolled slightly over 200 participants. Let us assume uniform annual rates of enrollment, uniform annual mortality, and follow-up of all participants to a common termination date. If an investigator were interested only in the experience of the participants over the initial 3 years after enrollment, assuming no further benefit from intervention after that time, then the single largest center would have required 21 years to recruit participants and 24 years to complete the study. Even if the investigator were interested simply in an equivalent number of person-years of intervention, regardless of the number of years a participant received the intervention, this one center would have taken approximately 12 years to complete the study.

   Given that medical advances are probably even more rapid today than when the Aspirin Myocardial Infarction Study was conducted, a 24-year study or even a 12-year study is impractical and may develop major problems. Changes in therapy and methodology during the years will make the study obsolete. Mortality from causes other than the one of interest may become more important in the later years of the study and dilute any effect of the intervention. It may not be reasonable to expect an intervention to continue to provide the same relative benefit over the course of many years. In addition, participants and investigators are likely to lose interest in the trial and may elect not to participate further. There is also a good chance that they may move from the area. Finally, answers from the trial, which might benefit other people, will be delayed for a generation. For these reasons, most investigators prefer to engage in studies of shorter duration.
2. A multicenter study may assure a more generalizable sample of the study or target population. Although no trial is completely representative, geography, race, socioeconomic status, and life style of participants may be more similar to the general population if participants are enrolled by many centers. These factors may be important in the ability to generalize the findings of the trial. Severity and sequelae of hypertension, for example, are seemingly race related. A study of hypertensive participants from either a totally black or totally white community are likely to yield findings that may not necessarily applicable to a more diverse population. Similarly, a study of pulmonary disease in an air-polluted industrial center might not give the same results as a study in a rural area.

3. A multicenter study enables investigators with similar interests and skills to work together on a common problem. Science and medicine, like many other disciplines, are competitive. Nevertheless, investigators may find that there are times when their own interests, as well as those of science, require them to cooperate. Thus, many scientists collaborate in order to solve particularly vexing clinical and public health problems and to advance knowledge in areas of common interest. A multicenter trial also gives capable, clinically oriented persons, who might otherwise not become involved in research activities, an opportunity to contribute to science. In the past, multicenter clinical trials typically involved only major academic centers. Now, many clinical practices based in the community successfully participate in trials.

Conduct of Multicenter Trials

One of the earlier multicenter clinical trials was the Coronary Drug Project [5]. This study provided an initial model for many of the techniques currently employed. Some techniques have been refined in subsequent trials. As in all active disciplines, concepts are frequently changing. Nonetheless, the following series of steps are one reasonable way to approach the planning and conduct of a multicenter trial. It consists of a distillation of experience from a number of these studies.

First, a planning committee should be established to be responsible for organizing and overseeing the various phases of the study (planning, participant recruitment, participant follow-up, phase out, data analysis, paper writing) and its various centers and committees. This group often consists of representatives from the sponsoring organization (e.g., government agencies, private research organizations, educational institutions, private industry), with input from appropriate consultants. Use of consultants who are expert in the field of study, in biostatistics, and in the management of multicenter clinical trials is encouraged. The planning committee needs to have authority in order to operate effectively and for the study to function efficiently.

Second, to determine the feasibility of a study, the planning committee should make a thorough search of the literature and review of other information. Sample size requirements should be calculated. Reasonable estimates must be made regarding control group event rate, anticipated effect of intervention, and participant adherence to therapy. The planning committee also has to evaluate key issues such as participant availability, availability of competent cooperating investigators, timeliness of the study, possible competing trials, regulatory requirements, and total cost. After such an assessment, is the trial worth pursuing? Are there sufficient preliminary indications that the intervention under investigation indeed might work? On the other hand, is there so much suggestive (though inconclusive) evidence in favor of the new intervention that it might be difficult ethically to allocate participants to a control group? Might such suggestive evidence seriously impede participant recruitment? Since planning for the study may take a year or more, feasibility needs constantly to be re-evaluated, even up to the time of the actual start of participant recruitment.

New or impending evidence may at any time cause cancellation, postponement, or redesign of the trial. In some instances, a pilot, or feasibility study is useful in answering specific questions important for the design and conduct of a full-scale trial.

Third, multicenter studies require not only clinical centers to recruit participants but also one or two coordinating centers to help design and manage the trial and to collect and analyze data from all other centers. There may be regional sites, academic centers, or contract research organizations (CROs) – also called clinical research organizations – that conduct site visits and receive data from the clinical centers. Additional centers are often needed to perform specialized activities such as key laboratory tests, imaging, and distributing study drugs. While the specialized centers may perform multiple services, it is usually not advisable to permit a clinical center to perform these services. If a specialized center and a clinical center are in the same institution, each should have a separate staff. Otherwise, unblinding and, therefore, bias could result. Even if unblinding or bias is avoided, there might be criticism that such a bias might have occurred and thus raise unnecessary questions about the entire clinical trial.

As reported by Croke [8], a major consideration when selecting clinical center investigators is availability of appropriate participants. Although this report is now old, the message remains relevant. The trial has to go where the participants are. Clearly, experience in clinical trials and scientific expertise are desirable features for investigators, but they are not crucial to overall success. Well-known scientists who add stature to a study are not always successful in collaborative ventures. The chief reason for this lack of success is often their inability to devote sufficient time to the trial. In a comprehensive study of factors associated with enrollment of eligible people with documented myocardial infarction, Shea et al. [9] found positive correlations with institutions in which patients were cared for by staff other than private attending physicians and with the presence of a committed nurse coordinator.

The selection of the coordinating center is of utmost importance. This is often a single entity, but sometimes, the coordinating center functions are split between two or more units; a clinical coordinating center, a data coordinating center, and, often, a separate data analysis center. The responsibilities described here apply to any of the models, but clearly communication becomes more of an issue when there are multiple units.

In addition to helping design the trial, the coordinating center, or combination of centers, is responsible for implementing the randomization scheme, for carrying out day-to-day trial activities, and for collecting, monitoring, editing, and analyzing data. The coordinating center, or, when there are two units, the clinical coordinating center/data management center needs to be in constant communication with all other centers. Its staff has to have expertise in areas such as biostatistics, computer technology, epidemiology, medicine, and management to respond expeditiously to daily problems that arise in a trial. These might range from simple questions, such as how to code a particular item on a questionnaire, to monitoring clinical site conduct. The single coordinating center, or the separate data analysis center, has responsibilities such as preparing data monitoring guidelines, conducting data analyses, and

developing or modifying statistical methods. The staffs at these centers must be experienced, capable, responsive, and dedicated to handle their workloads in a timely fashion. A trial can succeed despite inadequate performance of one or two clinical centers, but a poorly performing coordinating center or data management center can materially affect the success of a multicenter trial. In extreme cases, a coordinating center may have to be changed midway through the trial. This causes serious delay and logistical problems. Thus, proper selection of the coordinating center is extraordinarily important.

A key element in any coordinating or analysis center is not only the presence of integrity, but the appearance of integrity. Any suspicion of conflict of interest can damage the trial. It is for this reason that pharmaceutical firms who support trials sometimes use outside institutions or organizations as coordinating centers. Because the personnel in the centers control the data and the analyses, they should be seen to have no overriding interest in the outcome of a trial. Meinert [6] has described the functions of the coordinating center in detail. See also Fisher et al. for a description of the operations of an independent data analysis center [10]. As noted, certain functions in a multicenter trial are best carried out by properly selected special centers. The advantages of centrally performing laboratory tests, reading X-rays, evaluating pathology specimens, or coding electrocardiograms include unbiased assessment, standardization and reduced variability, ease of quality control, and high quality performance. The disadvantages of centralized determinations include the cost and time required for shipping, as well as the risk of losing study material. It is also obvious that the centers selected to perform specialized activities need expertise in their particular fields. Equally important is the capacity to handle the large workloads of a multicenter trial with research-level quality. Even with careful selection of these centers, backlogs of work are a frequent source of frustration during the course of a trial.

Fourth, it is preferable for the planning committee to provide prospective investigators with a fairly detailed outline of the key elements of the study design as early as possible. This results in more efficient initiation of the trial and allows each investigator to plan better his staffing and cost requirements. Rather than presenting a final protocol to the investigators, we recommended that all or selected representatives be given time to discuss and, if necessary, modify the trial design. This process allows them to contribute their own ideas, to have an opportunity to participate in the design of the trial, strengthening their commitment to it, and to become familiar with all aspects of the study. It may also improve the design. The investigators need a protocol that is acceptable to them and their colleagues at their local institution. This “buy-in” will improve participant recruitment, data collection, and final acceptance of the trial results. Depending on the complexity of the trial, several planning sessions prior to the start of participant recruitment may be needed for this process.

If there are a great many investigators and a number of difficult protocol decisions, it is useful during the planning stage to have specific groups or subsets of investigators to address these issues. Working groups can focus on individual problems and prepare reports for the total body of investigators. Of course, if the initial outline has been

well thought out and developed, few major design modifications will be necessary. Any design change needs to be carefully examined to ensure that the basic objectives and feasibility of the study are not threatened. This caveat applies particularly to modifications of participant eligibility criteria. Investigators are understandably concerned about their ability to enroll a sufficient number of participants. In an effort to make recruitment easier, they may favor less stringent eligibility criteria. Any such decisions need to be examined to ensure that they do not have an adverse impact on the objectives of the trial and on sample size requirements. The benefit of easier recruitment may be outweighed by the need for a larger sample size. Planning meetings also serve to make all investigators aware of the wide diversity of opinions. Inevitably, compromises consistent with good science must be reached on difficult issues, and some investigators may not be completely satisfied with all aspects of a trial. However, all are usually able to support the final design. All investigators in a cooperative trial must agree to follow the common study protocol.

Fifth, an organizational structure for the trial should be established with clear areas of responsibility and lines of authority. Many have been developed [11–15]. The one outlined below has stood the test of time.

Steering Committee. This committee provides scientific direction for the study at the operational level. Its membership may be made up of some or all of those who were on the planning committee (including sponsor representation) plus a subset of investigators participating in the trial. Depending on the length of the study, some key investigators may be permanent members of the Steering Committee to provide continuity. Others may be chosen or elected for shorter terms. Subcommittees are often established to consider on a study-wide level specific issues such as adherence, quality control, classification of response variables, and publication policies and review and then report to the Steering Committee.

It may also be important to authorize a small subgroup to make executive decisions between Steering Committee meetings. Most "housekeeping" tasks and day-to-day decisions can be more easily accomplished in this manner. A large committee, for example, is unable to monitor a trial on a daily basis, write memoranda, or prepare agendas. Since committee meetings can rarely be called at short notice, issues requiring rapid decisions must be addressed by an executive group. It is important, however, that major questions be discussed with the investigators.

Assembly of Investigators. This committee represents all of the centers participating in the trial. In small studies, this Assembly may be the same as the Steering Committee. In large studies, the Steering Committee would become too large to perform its duties effectively if it included all investigators. The purposes of Assembly meetings, which may be attended by other study personnel, are to allow for votes on major issues, to keep all investigators acquainted with the progress of the trial, and to provide an opportunity for staff training and education. Given the complexity of many trials, this last purpose is often the most important.

Subcommittees. Often, subcommittees of the Steering Committee are established. For example, there might be an Events Classification Subcommittee. Central evaluation of events, with the participant's identify and intervention group blinded,

helps to assure unbiased classification of reported events and to ensure consistent application of criteria for particular events. Other subcommittees might look for ways to improve participant accrual or adherence. In some trials, the subcommittee structure has become too complex and can lead to inefficiencies. Trials with few centers function best with a simple structure. If committees, subcommittees, and task forces multiply, the process of handling routine problems becomes difficult. Studies that involve multiple disciplines, especially need a carefully thought out organizational structure. Investigators from different fields tend to look at issues from various perspectives. Although this variety can be beneficial, under some circumstances it can obstruct the orderly conduct of a trial. Investigators may seek to increase their own areas of responsibility and, in the process, change the scope of the study. What starts out as a moderately complex trial can end up being an almost unmanageable undertaking.

Monitoring Committee. This scientific body, which goes by various names (see Chap. 16), should be independent of the investigators and any sponsor of the trial. It has as its primary roles the assurance, to the extent possible, of participant safety and study integrity. To accomplish those, it is charged with reviewing and approving the protocol, periodically monitoring baseline, toxicity, and response-variable data and evaluating center performance [16]. In the light of concerns about clinical-trial integrity [17–19], the independence of this group is especially important. It usually reports to either the study sponsor or the chairperson of the planning or steering committee. The coordinating or data analysis center should present tabulated and graphic data and appropriate analyses to the monitoring committee for review. The committee has the responsibility to recommend early termination in case of unanticipated toxicity, greater-than-expected benefit or high likelihood of indifferent results (see Chap. 16). Members of this committee should be knowledgeable in the field under study, in clinical trials methodology, and in biostatistics. An ethicist and/or a participant advocate may also be part of this group. The responsibilities of the monitoring committee to the participants, as well as to the integrity of the study, should be clearly established. These responsibilities for participant safety are particularly important in double-blind studies, since the individual investigators are unaware of the group assignments and which group is associated with various adverse events.

Sixth, despite special problems, multicenter trials should try to maintain standards of quality as high as those attainable in carefully conducted single-center trials. Therefore, strong emphasis should be placed on training and standardization. It is obviously extremely important that staff at all centers understand the protocol definitions, and how to complete forms and perform tests. Differences in performance among centers, as well as between individuals in a single center, are unavoidable. They can, however, be minimized by proper training, certification procedures, retesting, and when necessary, retraining of staff. These efforts need to be implemented before a trial gets underway. (See Chap. 11 for a discussion of quality control.) Not until the Steering Committee is satisfied that staff are capable of performing necessary procedures should a clinical center be allowed to begin enrolling participants. Meetings of the Assembly of Investigators are essential to the successful conduct of

the trial because they provide opportunities to discuss common problems and review proper ways to collect data and complete study forms.

Large simple trials [20] typically involve a large number of participating centers, many of which are nonacademic institutions. Education, training, and standardization may not get the same attention as in other trial models. Clinician-investigators need to understand the basic concepts and intent of clinical trials and how the rules of research – which may sometimes seem arbitrary – differ from the way they practice medicine. The reliance on hard endpoints such as all-cause mortality, and limited data collection in this kind of multicenter trial tends to reduce the need for elaborate quality control procedures.

Seventh, there needs to be close monitoring of the performance of all centers. Participant recruitment, quality of data collection and processing, quality of laboratory procedures, and adherence of participant to protocol should be evaluated frequently. Exactly how often is determined by the time span for which investigators are willing to allow errors or nonperformance to go undetected. Of course, cost and personnel considerations may dictate lesser frequency than desired. Tables 20.1–20.3 are typical of the kinds of information that investigators have used to compare clinical center performance.

Table 20.1 Average processing time for follow-up visit (FV) forms

	Previous 6 months		Present 6 months	
Clinic	No. of FV forms received	Days from visit to receipt of forms	No. of FV forms received	Days from visit to receipt of forms
A	292	25.8	290	8.7
B	157	22.9	117	29.0
C	210	16.0	198	16.2
D	174	11.6	173	10.4
E	182	8.3	185	12.7
Total	1,015	17.8	963	13.8

Table used in Aspirin Myocardial Infarction Study: Coordinating Center, University of Maryland

Table 20.2 Number of follow-up visit forms not received at coordinating center more than 1 month past the visit window, by clinic

Clinic	January 13, 1978	July 28, 1978
A	8	0
B	21	65
C	0	1
D	1	0
E	0	0
Total	30	66

Table used in Aspirin Myocardial Infarction Study: Coordinating Center, University of Maryland

Table 20.3 Percent of follow-up visit forms with one or more errors, by clinic and month of follow-up

Clinic	Feb.	Mar.	Apr.	May	June	July	Total	Total previous 6 months	Errors per form[a]	No. of forms processed
A	30.0	29.6	29.6	29.7	35.1	29.3	30.3	33.2	6.11	290
B	25.0	14.3	20.8	28.0	–	28.3	24.2	24.2	6.66	117
C	0.0	14.1	3.4	8.1	27.3	13.8	12.1	16.2	5.21	198
D	22.2	16.7	6.3	20.9	9.7	26.3	16.2	17.8	6.21	173
E	4.8	10.3	19.6	13.6	21.2	20.8	15.7	18.7	4.38	185
Total	17.0	18.5	17.0	20.4	20.9	23.8	20.6	23.1	5.68	963

Table used in Aspirin Myocardial Infarction Study: Coordinating Center, University of Maryland
[a]Errors per form are calculated by dividing the number of errors by the number of forms failing edit

Table 20.1 compares the average time it takes for each clinical center to complete and submit forms to the coordinating center. Center B in the present 6 months stands out as doing poorly and it has become worse when compared with previous performance. Table 20.2 shows that center B also has a large number of unsubmitted forms. As seen in Table 20.3, clinical centers A and B are submitting many forms with errors. It is useful to identify those centers which are performing below average. Often, specific problems can be identified and corrected. For example, in one study, evidence of left ventricular hypertrophy was identified much more often in the electrocardiograms from one center than from the other centers. Only after looking into the reasons for this was it discovered that the internal standard on that clinical center's electrocardiograph machine was incorrectly calibrated.

Many industry-sponsored multicenter trials that employ CROs conduct extensive auditing and quality assurance. This is quite costly and how much benefit it provides has been questioned [21]. See Chap. 11 for further discussion of this topic.

In all clinical trials, recruitment of participants is difficult. In a cooperative clinical trial, however, there is an opportunity for some clinical centers to compensate for the inadequate performance of other centers by exceeding their predetermined recruitment goals. The clinical centers should understand that, while friendly competition keeps everybody working, the real goal is overall success, and what some centers cannot do, another perhaps can. Therefore, it is important to encourage the good centers to recruit as many participants as possible. There may be a limit, however, if one center, region, or country (in the case of international trials) starts to dominate enrollment. At some point, recruitment might need to be capped if the study is to be seen as truly multicenter.

Eighth, publication, presentation, and authorship policies should be agreed upon in advance. Authorship becomes a critical issue when there are multiple investigators, many of whose academic careers depend on publications. Unfortunately, there is no completely satisfactory way to recognize the contribution of each investigator. A common compromise is to put the study name immediately under the paper title and to acknowledge the writers of the paper, either in a footnote or under the title, next to the study name. All key investigators are then listed at the end of the paper. The policy may also vary according to the type of paper (main or subsidiary).

The group authorship of manuscripts from multicenter trials was challenged by some medical journals and defended by others [22–24]. It remains common, but typically with an identified writing committee to take responsibility (see Chap. 19).

Involvement of the sponsor as an author of the main manuscripts from a major trial can be contentious, especially if it is a commercial firm that stands to benefit from a favorable presentation of the trial results. Most sponsors accept a hands-off policy and leave it to the investigators to write the scientific papers. Typically, they are given 1 month to preview a manuscript, particularly for patent or regulatory issues. This review should not unnecessarily delay the publication of the trial results. Regrettably, there are examples of interference that is in conflict with academic freedom.

In one four-center trial, the investigators at one of the centers reported their own findings before the total group had an opportunity to do so [25, 26]. Such an action is not compatible with a collaborative effort. It undermines the goal of a multicenter trial of having enough participants to answer a question and, perhaps more importantly, the trust among investigators. Academic institutions have taken a strong stand against this principle of collaboration and in defense of academic freedom for each investigator. However, we believe that those unwilling to abide by the rule for common authorship should not participate in collaborative studies.

Advance planning of authorship policy may eliminate subsequent misunderstandings. However, fair recognition of junior staff will always be difficult [27]. Study leadership often gets credit and recognition for work done largely by people whose contributions may remain unknown to the scientific community. One way to alleviate this problem is to appoint as many capable junior staff as possible to subcommittees. Such staff should also be encouraged to develop studies ancillary to the main trial. This approach will enable them to claim authorship for their own work while using the basic structure of the trial to get access to participants and supporting data. Such ancillary studies may be performed on only a subgroup of participants and may not necessarily be related to the trial as a whole. Care must be taken to ensure that they do not interfere with the main effort, either through unblinding, by harming the participants or by causing the participants to leave the trial. Sackett and Naylor discuss the issues for and against allowing publication of ancillary studies before the main trial is completed [28].

Globalization of Trials

As noted earlier, many multicenter clinical trials are international. The reasons are several. One, it provides a greater number of potential participants, allowing for quicker accrual. Two, the broader populations may allow for wider generalization of results. It is not simply people from one country with one medical care system who are enrolled. The data from the trial apply to many sorts of people with very different medical systems. Three, it may be easier and less expensive to screen people in some regions.

There are, however, limitations and concerns. As discussed in Chap. 2, the ethics of enrolling participants from underdeveloped countries or areas can be problematic.

It is unethical to enter people into a trial simply to save money, or because the regulatory oversight is less rigorous, when there is little likelihood that the population will benefit from or have access to the trial intervention. Logistics of implementing an international trial may be daunting. In addition to multi-language communication, there is the issue of translating forms. Not all forms, particularly those that have been validated in certain groups, may be usable in very different communities and cultures. Transporting drugs and other materials across borders may not be simple. And, of course, each country will have its own regulatory structure that must be negotiated. Study leadership and proper representation on Steering Committees and Monitoring Committees must be carefully worked out.

Interpretation of results may be questioned. Are the overall results relevant to all countries? Does the culture, social structure, or medical care system (including concomitant medications and other treatment) affect the outcome? Does each trial participant need minimal standard background care? If so, this must be specified in advance, in the protocol. An example of a trial that examined effect by geography is the Platelet IIb/IIIa in Unstable angina: Receptor Suppression Using Integrilin Therapy, or PURSUIT trial [29]. Relative reductions in the primary response variable (a combination of death or myocardial infarction) varied among geographic regions. After adjustment for baseline factors, much of this difference was eliminated. Another cardiovascular trial, the Metoprolol CR/XL Randomized Intervention Trial in Chronic Heart Failure (MERIT-HF) observed a hazard ratio of close to 1 for mortality (not the primary response variable) among US sites, whereas the hazard ratio was 0.55 for the European sites. Although it is unlikely that these subgroup findings in a secondary (though obviously important) outcome were due to anything but chance [30], some asked if the European results could be applied to the USA. In these examples, chance and other explanations (baseline difference) were likely explanations. In other trials, observed differences might be harder to dismiss, and investigators need to consider, in advance, whether combining results from geographically and culturally different sites is appropriate. O'Shea and Califf [31] discuss difference in outcome of cardiovascular trials among countries. Vickers et al. [32] found that some countries tended to produce results more favorable to the new intervention than other countries, though publication bias was the likely reason.

General Comments

Even if investigators think they have identified all potential difficulties and have taken care to prevent them, new problems will always arise. This is particularly true of multinational trials because of their size, complexity and large number of investigators with diverse backgrounds, interests, medical practices, and language. To forestall and minimize problems, the need for adequate study-wide communication must be stressed. If communication among the various components of the study lapses, or is vague, the trial can rapidly deteriorate. It is the responsibility of coordinating centers to keep in frequent contact (by telephone, e-mail, letter, and visit) with all the other centers. This contact needs to be initiated by both coordinating centers and

other centers. The study leaders also need to maintain contact with the various centers and committees, closely monitoring the conduct of the trial.

Much communication is done via the internet. With proper attention to maintenance of confidentiality, this has proven to be an effective and relatively inexpensive way for investigators to interact, particularly when there are large time differences among centers (see Chap. 11).

Difficulty in getting groups of clinicians to work together using a common protocol has been reported [33]. A group of experienced Italian scientists tried to engage general practitioners in a drug trial of people with isolated systolic hypertension. The well-established principles for organizing a collaborative study were followed. The practitioners were informed and trained and were also given the opportunity to comment on the study protocol. Unfortunately, only 88 of the 806 general practitioners who had agreed to participate eventually started recruitment. Sixty-three enrolled at least one participant. Due to poor cooperation, the study was stopped after completion of its feasibility phase. A major problem was the practitioners' difficulties in withdrawing drug therapy from hypertensive patients. "As noted by the authors of the article, the change from the role of confident and reassuring prescriber to an attitude of uncertainty (which attracted consensus in the preparatory meetings) raised instinctive resistance in practice, leading to the withdrawal of the general practitioner rather than the patient's treatment."

Cost is always a concern in multicenter trials. These studies are generally expensive due to their complexity, size, and (sometimes) elaborate committee structure. Expense can be minimized by asking only pertinent questions on forms, by reducing the number of laboratory tests, and by performing only necessary quality monitoring; in essence, by simplifying data collection [34]. Accomplishing these economies demands constant attention, particularly during the planning phase. The investigators in a multicenter trial traditionally represent diverse interests. Given the opportunity, most of them would pursue these interests. Few would like to miss an opportunity to add to the main trial questions or examinations of particular importance to them. These additions are often important scientifically. However, it is easy for a trial to become overbuilt and get out of control. It is usually a good policy to restrict additions to the basic study protocol. Special caution should be taken when the argument for inclusion is, "it would be interesting to know." The purpose of every procedure and item on the study forms should be clearly defined in advance and a test hypothesis formulated, if possible. Most late-phase trials collect more data than is ever used for primary or even secondary publications. Certain questions can be answered by using less than the total number of participants. Other questions require the whole group. Still others cannot be answered, even if all participants are included. Therefore, in each instance, thought should be given to sample size and power calculations.

References

1. Fleiss JL. Multicentre clinical trials: Bradford Hill's contributions and some subsequent developments. *Stat Med* 1982;1:353–359.
2. Greenberg BG. Conduct of cooperative field and clinical trials. *Am Stat* 1959;13:13–28.

3. Klimt CR. Principles of multi-center clinical studies. In Boissel JP, Klimt CR. *Multi-center Controlled Trials. Principles and Problems.* 1979; INSERM, Paris.
4. Levin WC, Fink DJ, Porter S, et al. Cooperative clinical investigation: a modality of medical science. *JAMA* 1974;227:1295–1296.
5. Coronary Drug Project Research Group. PL Canner (ed.). The Coronary Drug Project: methods and lessons of a multicenter clinical trial. *Control Clin Trials* 1983;4:273–541.
6. Meinert CL. *Clinical Trials. Design, Conduct and Analysis*. New York: Oxford University Press, 1986.
7. Aspirin Myocardial Infarction Study Research Group. A randomized, controlled trial of aspirin in persons recovered from myocardial infarction. *JAMA* 1980;243:661–669.
8. Croke G. Recruitment for the National Cooperative Gallstone Study. *Clin Pharmacol Ther* 1979;25:691–694.
9. Shea S, Bigger JT, Campion J, et al. Enrollment in clinical trials: institutional factors affecting enrollment in the Cardiac Arrhythmia Suppression Trial (CAST). *Control Clin Trials* 1992;13:466–486.
10. Fisher MR, Roecker EB, DeMets DL. The role of an independent statistical analysis center in the industry-modified National Institutes of Health model. *Drug Inf J* 2001;35:115–129.
11. Byington RP, for the Beta-Blocker Heart Attack Trial Research Group. Beta-Blocker Heart Attack Trial: design, methods, and baseline results. *Control Clin Trials* 1984;5:382–437.
12. Meinert CL. Organization of multicenter clinical trials. *Control Clin Trials* 1981;1:305–312.
13. Lachin JM, Marks JW, Schoenfield LJ, et al. Design and methodological considerations in the National Cooperative Gallstone Study: a multicenter clinical trial. *Control Clin Trials* 1981;2:177–229.
14. Carbone PP. Organization of clinical oncology in the U.S.A.: role of cancer centers, cooperative groups and community hospitals. *Eur J Cancer Clin Oncol* 1985;21:119–154.
15. Carbone PP, Tormey DC. Organizing multicenter trials: lessons from the cooperative oncology groups. *Prev Med* 1991;20:162–169.
16. Friedman L, DeMets D. The data monitoring committee: how it operates and why. *IRB* 1981;3:6–8.
17. Fleming TR, DeMets DL. Monitoring of clinical trials: issues and recommendations. *Control Clin Trials* 1993;14:183–197.
18. Angell M, Kassirer JP. Setting the research straight in the breast-cancer trials. *N Engl J Med* 1994;330:1448–1450.
19. Cohen J. Clinical trial monitoring: hit or miss? *Science* 1994;264:1534–1537.
20. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med* 1984;3:409–422.
21. Eisenstein EL, Lemons PW 2nd, Tardiff BE, et al. Reducing the costs of phase III cardiovascular clinical trials. *Am Heart J* 2005;149:482–488.
22. Kassirer JP, Angell M. On authorship and acknowledgments. *N Engl J Med* 1991;325:1510–1512.
23. Goldberg MF. Changes in the archives. *Arch Ophthalmol* 1993;111:39–40.
24. Meinert CL. In defense of the corporate author for multicenter trials. *Control Clin Trials* 1993;14:255–260.
25. Winston DJ, Ho WG, Gale RP. Prophylactic granulocyte transfusions during chemotherapy of acute nonlymphocytic leukemia. *Ann Intern Med* 1981;94:616–622.
26. Strauss RG, Connett JE, Gale RP, et al. A controlled trial of prophylactic granulocyte transfusions during initial induction chemotherapy for acute myelogenous leukemia. *N Engl J Med* 1981;305:597–603.
27. Remington RD. Problems of university-based scientists associated with clinical trials. *Clin Pharmacol Ther* 1979;25:662–665.
28. Sackett DL, Naylor CD. Should there be early publication of ancillary studies prior to the first primary report of an unblinded randomized clinical trial? *J Clin Epidemiol* 1993;46:395–402.
29. Akkerhuis KM, Deckers JW, Boersma E, et al, for the PURSUIT Investigators. Geographic variability in outcomes within an international trial of glycoprotein IIb/IIIa inhibition in patients with acute coronary syndromes: results from PURSUIT. *Eur Heart J* 2000;21:371–381.

30. Wedel H, DeMets D, Deedwania P, et al, on behalf of the MERIT-HF Study Group. Challenges of subgroup analyses in multinational clinical trials. Experiences from the MERIT-HF trial. *Am Heart J* 2001;143:502–511.
31. O'Shea JC, Califf RM. International differences in treatment effects in cardiovascular clinical trials. *Am Heart J* 2001;141:875–889.
32. Vickers A, Goyal N, Harland R, Rees R. Do certain countries produce only positive results? A systematic review of controlled trials. *Control Clin Trials* 1998;19:159–166.
33. Tognoni G, Alli C, Avanzini F, et al. Randomised clinical trials in general practice: lessons from a failure. *Br Med J* 1991;303:969–971.
34. Eisenstein EL, Collins R, Cracknell BS, et al. Sensible approaches for reducing clinical trial costs. *Clin Trials* 2008;5:75–84.

Index

A

accidental bias, 98
adaptive design, 90–91
  response adaptive, 90
  sample size, 90–91
  trend adaptive, 90–91
adaptive randomization, 105–109
adherence, 61–62, 251–268, 418
  considerations before enrollment, 253–258
  dealing with low adherence, 265–266
  definition, 251
  design factors, 253–255
  maintaining, 258–262
  monitoring, 262–265
  run-in, 254–255
  withdrawal, 253
adverse effects, 233–237
adverse events, 60, 215–229
  analysis, 223–224
  ascertainment, 220–221
  classification, 219
  CONSORT statement, 226
  definition, 218–219
  dimensions, 221
  identification, 227–228
  meta-analysis, 228–229
  regulatory, 226–227
  reporting, 224–227
  serious, 216–218
  unexpected, 218
alpha spending functions for sequential monitoring, 318–321
alternative hypothesis, 135
analysis, 345–390
  intention-to-treat, 3, 346–355, 412
  on-treatment, 350–355
  per protocol, 412
  trend adaptive designs, 389
ascertainment
  post-study, 402
asymmetric group sequential boundaries, 324–325
audits, 210–212

B

baseline adaptive randomization, 105–108
baseline data, 169–180
  comparability, 170–171, 179–180
  definition, 169
  measurement, 173
  testing for imbalance, 180
  uses, 169–174
Bayesian methods, 5–7
Bayesian methods for data monitoring, 330–331
Belmont report, 24
bias, 119–121, 206
biased coin randomization, 105–106
biomarkers, 47–50
blindness, 1, 119–131
  assessment, 129–131
  double-blind design, 122–123
  “double-dummy,” 127
  protecting, 124–129
  reporting, 129–131, 417–418
  single-blind design, 120–122
  triple-blind design, 123–124
  unblinded (open) design, 119–120
  unblinding, 124–125, 127–129
blocked randomization, 100–102
Bonferoni procedure for multiple testing, 377

C

case-control studies, 2, 71
censoring, 272–274

classical sequential monitoring, 314–315
clinical equipoise, 22, 295
close-out, 399–409
  transfer of care, 402–403
  visit, 400
Common Rule, 24
comparison of multiple variables, 377–378
comparison of survival curves, 279–285
  Cox Proportional Hazards Model, 289
  Mantel-Haenszel, 279–284
  medians, 279–280
  point by point, 279
  Weighted Rank tests, 286
  Wilcoxon-Gehan, 283
compensatory treatment, 121–122
competing event issues, 61, 362
compliance (see adherence)
composite event issues, 44–45, 363
concomitant treatment, 121–122, 418
concurrent control studies, 69–73
  non-randomized, 72–73
conditional power monitoring procedures, 325–330
conflict of interest, 23–24, 414
CONSORT statement, 130–131
continuous response, 60, 135, 147–150
control group, 2, 22–23
  concurrent, non-randomized, 72–73
  historical, 73–79
  placebo, 22–23
  randomized, 69–72
covariate adjustment in analysis, 364–370
  using baseline data, 368–370
  using surrogates, 365–367
covariate-intervention interaction, 368–370
cross-over, 79–80, 252–253
Cutler-Ederer, 271, 278

D
data analysis issues, 345–390
data and safety monitoring board (see data monitoring committee)
databases, 73–79
data collection, 199–212
  data entry, 206–207
  enhancing quality, 203–207
  essential data, 200–201
  guidelines, 199
  problems, 201–202
  verification, 403–404
data missing completely at random, 355
data monitoring committee, 295–299
  decision process, 298, 301–310
  meetings, 297–299
  membership, 296
data sharing, 26, 30
Declaration of Helsinki, 24
dichotomous response, 135, 139–146
dose-escalation studies, 5
double-blind design (see blindness)
drop-in, 252–253
drop-out, 252–253
drugs
  coding, 127
  matching, 125–127
  quality control, 209–210

E
effectiveness trial, 3
efficacy trial, 3
equivalence study, 86–90
ethics, 19–32, 123–124
  clinical equipoise, 22, 295
  conflict of interest, 23–24, 32
  data monitoring, 29
  developing countries, 27–28
  emergency settings, 26–27
  financial incentives, 28
  guidelines, 25
  informed consent, 24–27
  privacy and confidentiality, 30–31
  publication bias, 31–32
  surrogate consent, 27
Ethics Review Committee, 19
excluding participants from analysis, 346–355
  ineligibility, 347–350
  missing data, 355–362
  nonadherence, 350–355
expected number of events, 154–155
exponential distribution for survival outcome, 152, 271

F
factorial design, 82–83
falsification of data, 31
fixed allocation, 98–105
follow-up, poststudy, 407–409
forms
  development, 203–204
  monitoring, 208
  pretesting, 205–206

G
ghost authorship, 32, 414
Greenberg Report, 296
Greenwood's variance formula, 277

group allocation design, 83–84
group sequential methods for monitoring, 315–325
group sequential monitoring applications, 321–324
guest authorship, 32, 414

H
Haybittle-Peto group sequential method, 316
health-related quality of life, 233–247
  administration, 242–243
  clinical significance, 245–246
  components, 233
  definition, 233
  design, 239–240
  interpretation, 244–247
  methodology, 239–244
  outcomes, 237–239
  pretesting, 239
  primary dimensions, 234–236
  selection of instruments, 241–242
  uses, 237–239
  utility measures, 246–247
historical control studies, 73–79
  limitations, 74–78
  role, 78–79
  strengths, 73–74
Hochberg procedure for multiple testing, 378
Holm procedure for multiple testing, 377
hybrid design, 84

I
imbalance, 122
imputation methods for missing data, 355, 356
industry modified NIH clinical trials model, 294
informative censoring or missing data, 357–361
informed consent, 24–27
Institutional Review Board (IRB), 19, 24–27
intention-to-treat (ITT) principle, 3, 346–355, 412
interim events, 176–177
intervention, 2, 42–43
IVRS randomization, 110
IWRS randomization, 110

K
Kaplan-Meier, 271, 274–278

L
Lan-DeMets alpha spending functions, 319–321
large, simple trials, 41, 57, 84–86
last observation carried forward, 356
levels of significance, 133

M
Mantel Haenszel statistic, 280–286, 389
manual of procedures, 203–204
margin of indifference, 42, 87, 89, 379–382
margin of noninferiority (see margin of indifference)
maximally tolerated dose, 5
meta-analysis, 171–172, 382–389
  methods, 386–389
minimization allocation, 107
monitoring committee, 123–124, 295–299
monitoring response variables, 293–334
multicenter trials, 427–438
  conduct, 429–436
  globalization, 436–437
  organization, 432–435
  reasons for, 428
  recruitment, 190

N
natural history, 40–41, 173–174
NIH clinical trials model, 293
nonadherence
  analysis, 350–355
  impact on sample size, 145–146, 252
noninferiority studies, 41–42, 86–90, 121
  assay sensitivity, 86, 381
  constancy assumption, 88
  control group, 86, 89
  efficacy imputation, 89
  interpretation, 87–90
  margin of indifference, 42, 87, 89
noninferiority trial analysis, 379–382
noninferiority trial assay sensitivity, 381
noninferiority trial margins, 42, 87, 379–382
null hypothesis, 135
Nuremberg Code, 24

O
O'Brien-Fleming group sequential method, 316–321
observational studies, 11
one-sided *vs*. two-sided hypothesis, 137
open trial (see blindness)

P
per protocol analysis, 412
pharmacogenetics, 172–173
phases of trials, 3–8
  early phase, 2, 4, 21
  late phase, 4–8
  phase 1, 2, 4–6
  phase 2, 2, 6–7
  phase 3, 7–8
  phase 4, 7–8
pilot study, 188, 191
placebo, 123, 125, 127–129
placebo control, 22–23
play-the-winner, 108
Pocock group sequential method, 316–317
Pocock-Simon randomization, 107
pragmatic trial, 3, 63, 84–86
privacy, 190, 401
procedures
  monitoring, 208–209
protocol, 12–14
publication bias, 31–32, 416–417

Q
quality control, 203–212
  certification, 204–205
  monitoring, 207–212
quality of life (see health-related quality of life)
questions, 37–51
  adverse event, 39
  ancillary, 40
  natural history, 40–41
  primary, 38
  secondary, 38–39

R
randomization, 1, 22, 97–113
randomization mechanics, 109–111
randomized control trials, 69–72
recruitment, 64–65, 183–197
  approaches to lagging recruitment, 194–197
  conduct, 190–192
  contingency plan, 187–188
  ethics, 28
  logs, 190
  monitoring, 192–194
  planning, 186–188
  problems, 184–186
  reasons for participating, 184
  “recycling” subjects, 196
  selection of study sample, 184
  sources, 188–190
registration of trials, 32, 417
regression toward the mean, 175–176
repeated confidence intervals, 322–323
repeated testing for significance, 299–301
reporting of results, 411–422
  analysis, 419–430
  guidelines, 413–415
  interpretation, 415
  publication bias, 31–32, 416–417
response adaptive randomization, 108
response variable, 43–45
  assessment, 46–47, 401–402
  combined (composite), 44–45
  monitoring, 293–334
  surrogate, 47–50
results
  dissemination, 405–407

S
safety, 215–217
sample size, 133–162
  cluster designs, 157–159
  continuous outcome for repeated measures, 150–152
  continuous outcome for two independent samples, 147–148
  continuous outcome paired data, 148–150
  dichotomous outcome for paired data, 144–145
  dichotomous outcome for two independent samples, 139–144
  multiple outcomes, 161–162
  nonadherence adjustment, 145–146, 150
  noninferiority designs, 155–157
  parameter estimation, 159–161
  time to failure outcome, 152–155
screening, 174–175, 191
selection bias, 97
simple randomization, 99–100
single-blind design (see blindness)
staggered entry, 271
statistical methods for data monitoring, 313–332
statistical power, 136
storage of study material, 404–405
stratification, 171
stratified randomization, 102–105
study population, 55–65
  definition, 55
  eligibility criteria, 57–62, 64–65
  external validity, 62

generalization, 62–64
heterogeneity, 59
homogeneity, 58, 64
large, simple trials, 57
recruitment, 64–65
representativeness, 63
selection, 58–60
subgroup analysis, 371–376
definition, 371
subgrouping, 171–172
subgroups, 38–39
superiority trials, 21, 41–42
surrogate consent, 27
surrogate response variable, 47–50
survival analysis methods, 269–290
survival curve estimation, 270–278

T
termination of trial
planning, 399–400
procedures, 399–403
therapeutic misconception, 26
time to failure response, 135, 152–155
training, 204–205, 239, 243
trial extension, 310–313
trial termination examples, 301–310
triple-blind design (see blindness)
two-armed bandit randomization, 108
type I error, 135
type II error, 136

U
unblinded design (see blindness)
unblinding, 124–125, 127–129, 403
urn design randomization, 106

V
variability, 175–176, 201–202, 206–207

W
withdrawal of subjects, 226, 346–347
withdrawal studies, 81