Praise for previous editions

“A concise, easily readable, and thorough summary...ALL medical oncology, radiation oncology, surgical oncology, and clinical research nurse academic training programs should provide this important text to trainees on Day 1.”
—Charles R. Thomas Jr., MD, University of Texas Health Science Center at San Antonio

“With over 60 years combined experience, the authors are ideally positioned to discuss the various statistical issues apparent in clinical trials, identifying alternative solutions, providing logical arguments for and against the various solutions. This book is also recommended for statisticians actively involved in the design, conduct, and analysis of clinical trial data (not only cancer clinical trials).”
—Journal of Biopharmaceutical Statistics

The third edition of the bestselling Clinical Trials in Oncology provides a concise, nontechnical, and thoroughly up-to-date review of methods and issues related to cancer clinical trials. The authors emphasize the importance of proper study design, analysis, and data management and identify the pitfalls inherent in these processes. In addition, the book has been restructured to have separate chapters and expanded discussions on general clinical trials issues, and issues specific to Phases I, II, and III. New sections cover innovations in Phase I designs, randomized Phase II designs, and overcoming the challenges of array data.

Although this book focuses on cancer trials, the same issues and concepts are important in any clinical setting. As always, the authors use clear, lucid prose and a multitude of real-world examples to convey the principles of successful trials without the need for a strong statistics or mathematics background. Armed with Clinical Trials in Oncology, Third Edition, clinicians and statisticians can avoid the many hazards that can jeopardize the success of a trial.
Clinical Trials in Oncology
Third Edition
CHAPMAN & HALL/CRC
Interdisciplinary Statistics Series
Series editors: N. Keiding, B.J.T. Morgan, C.K. Wikle, P. van der Heijden

Published titles

AN INVARIANT APPROACH TO STATISTICAL ANALYSIS OF SHAPES
S. Lele and J. Richtsmeier

ASTROSTATISTICS
G. Babu and E. Feigelson

BAYESIAN ANALYSIS FOR POPULATION ECOLOGY
Ruth King, Byron J.T. Morgan, Olivier Gimenez, and Stephen P. Brooks

BAYESIAN DISEASE MAPPING: HIERARCHICAL MODELING IN SPATIAL EPIDEMIOLOGY
Andrew B. Lawson

BIOEQUIVALENCE AND STATISTICS IN CLINICAL PHARMACOLOGY
S. Patterson and B. Jones

CLINICAL TRIALS IN ONCOLOGY, THIRD EDITION
S. Green, J. Benedetti, A. Smith, and J. Crowley

CLUSTER RANDOMISED TRIALS
R.J. Hayes and L.H. Moulton

CORRESPONDENCE ANALYSIS IN PRACTICE, SECOND EDITION
M. Greenacre

DESIGN AND ANALYSIS OF QUALITY OF LIFE STUDIES IN CLINICAL TRIALS, SECOND EDITION
D.L. Fairclough

DYNAMICAL SEARCH
L. Pronzato, H. Wynn, and A. Zhigljavsky

FLEXIBLE IMPUTATION OF MISSING DATA
S. van Buuren

GENERALIZED LATENT VARIABLE MODELING: MULTILEVEL, LONGITUDINAL, AND STRUCTURAL EQUATION MODELS
A. Skrondal and S. Rabe-Hesketh

GRAPHICAL ANALYSIS OF MULTI-RESPONSE DATA
K. Basford and J. Tukey

MARKOV CHAIN MONTE CARLO IN PRACTICE
W. Gilks, S. Richardson, and D. Spiegelhalter
Published titles

INTRODUCTION TO
COMPUTATIONAL BIOLOGY:
MAPS, SEQUENCES, AND GENOMES
M. Waterman

MEASUREMENT ERROR AND
MISCLASSIFICATION IN STATISTICS
AND EPIDEMIOLOGY: IMPACTS AND
BAYESIAN ADJUSTMENTS
P. Gustafson

MEASUREMENT ERROR:
MODELS, METHODS, AND APPLICATIONS
J. P. Buonaccorsi

META-ANALYSIS OF BINARY DATA
USING PROFILE LIKELIHOOD
D. Böhning, R. Kuhnert,
and S. Rattanasiri

STATISTICAL ANALYSIS OF GENE
EXPRESSION MICROARRAY DATA
T. Speed

STATISTICAL AND COMPUTATIONAL
PHARMACOGENOMICS
R. Wu and M. Lin

STATISTICS IN MUSICOLOGY
J. Beran

STATISTICS OF MEDICAL IMAGING
T. Lei

STATISTICAL CONCEPTS
AND APPLICATIONS IN
CLINICAL MEDICINE
J. Aitchison, J.W. Kay,
and I.J. Lauder

STATISTICAL AND PROBABILISTIC
METHODS IN ACTUARIAL SCIENCE
P.J. Boland

STATISTICAL DETECTION AND
SURVEILLANCE OF GEOGRAPHIC
CLUSTERS
P. Rogerson and I. Yamada

STATISTICS FOR ENVIRONMENTAL
BIOLOGY AND TOXICOLOGY
A. Bailer and W. Piegorsch

STATISTICS FOR FISSION
TRACK ANALYSIS
R.F. Galbraith

VISUALIZING DATA PATTERNS
WITH MICROMAPS
D.B. Carr and L.W. Pickle
Contents

1. Introduction .......................................................... 1
  1.1 A Brief History of Clinical Trials ............................... 1
  1.2 The Southwest Oncology Group (SWOG) ....................... 5
  1.3 The Reason for This Book .................................. 6

2. Statistical Concepts .................................................. 7
  2.1 Introduction ...................................................... 7
  2.2 The Single-Arm Phase II Trial—Estimation .................. 14
  2.3 The Randomized Phase III Trial—Hypothesis Testing ........ 18
    2.3.1 Response as the Outcome ............................... 18
    2.3.2 Survival as the Outcome ............................... 23
  2.4 The Proportional Hazards Model ............................... 30
  2.5 Sample Size Calculations .................................. 32
  2.6 Concluding Remarks ........................................ 33

3. The Design of Clinical Trials ...................................... 35
  3.1 Objectives ..................................................... 36
  3.2 Eligibility ..................................................... 36
  3.3 Treatment Arms ................................................ 37
    3.3.1 Single Arm ............................................ 37
    3.3.2 Two or More Treatment Arms ......................... 38
  3.4 Randomized Treatment Assignment ............................ 39
    3.4.1 Blinding ............................................... 39
  3.5 Endpoints ..................................................... 41
    3.5.1 Survival .............................................. 41
    3.5.2 Progression-Free Survival (PFS) ....................... 42
    3.5.3 Response .............................................. 43
    3.5.4 Toxicity Criteria ..................................... 44
    3.5.5 Quality of Life ....................................... 45
  3.6 Differences to be Detected or Precision of Estimates
    and Other Assumptions ....................................... 46
  3.7 Use of Independent Data Monitoring Committees ............ 47
    3.7.1 Composition ........................................... 49
    3.7.2 Concluding Remarks on Monitoring Committees ........ 51
  3.8 Ethical Considerations ....................................... 52
  3.9 Conclusion .................................................... 55
4. Phase I and Phase I/II Trials ............................................ 57
   4.1 Phase I Trials .................................................. 57
   4.1.1 The Traditional 3 + 3 Design ............................... 57
   4.1.2 Improving Phase I Designs .................................. 59
       4.1.2.1 Start at a Higher Dose? .......................... 60
       4.1.2.2 Modify the Traditional Algorithm? ............... 60
       4.1.2.3 Use Model-Based Designs? ....................... 63
       4.1.2.4 Alternative Approaches for Biologic Agents?  66
   4.1.3 Phase I Conclusion ......................................... 66
   4.2 Phase I/II Designs ............................................. 67
       4.2.1 Combining Phase I and Phase II ........................ 67
       4.2.2 Phase I/II ............................................... 68
       4.2.3 Phase I/II Conclusion .................................. 70

5. Phase II Trials .......................................................... 73
   5.1 Single-Arm Phase II Designs .................................... 73
       5.1.1 The Standard SWOG Phase II Design ................... 74
       5.1.2 Other Single-Arm Phase II Designs .................... 76
       5.1.3 Alternative Endpoints .................................. 77
       5.1.4 Single-Arm Pilot Designs ............................... 77
   5.2 Multi-Arm Phase II Trials ..................................... 77
       5.2.1 Nonrandomized Phase II Designs with a Control .... 77
       5.2.2 Randomized Phase II Designs with a Control ....... 78
       5.2.3 Randomized Selection Designs ........................ 79
       5.2.4 Other Randomized Designs ............................. 81
   5.3 Other Phase II Designs ......................................... 82
       5.3.1 Multiple Endpoint Designs ............................ 82
       5.3.2 Multi-Strata Trials .................................... 83
   5.4 Randomized versus Single Arm: The Pros and Cons ........ 83
   5.5 Conclusion ..................................................... 86

6. Phase III Trials .......................................................... 87
   6.1 Randomization .................................................. 87
       6.1.1 Stratification Factors .................................. 88
       6.1.2 Timing of Randomization ............................... 90
   6.2 Other Design Considerations .................................. 92
       6.2.1 One-Sided or Two-Sided Tests ........................ 92
       6.2.2 Significance Level, Power, and Sample Size .......... 92
       6.2.3 Multiple Endpoints .................................... 94
   6.3 Equivalence or Noninferiority Trials ......................... 95
       6.3.1 Designing an Equivalence or Noninferiority Trial .. 97
   6.4 Designs for Targeted Agents ................................... 98
   6.5 Multi-Arm Trials ............................................... 102
       6.5.1 Types of Multi-Arm Trials ............................. 102
       6.5.2 Significance Level .................................... 103
       6.5.3 Power .................................................. 104
## Contents

6.5.4 Interaction ............................................ 106  
6.5.5 Other Model Assumptions .......................... 110  
6.5.6 Sequential Randomization ......................... 110  
6.5.7 Concluding Remarks on Multi-Arm Trials ......... 112  
6.6 Interim Analyses ........................................... 113  
6.6.1 Examples of Interim Analyses ....................... 117  
6.6.1.1 Stopping Early for Positive Results .......... 118  
6.6.1.2 Stopping Early for Negative Results ......... 118  
6.6.1.3 Stopping an Equivalence Trial Early for Positive Results .......... 121  
6.6.1.4 Stopping Based on Toxicity and Lack of Compliance ................. 123  
6.6.1.5 Emergency Stopping Based on Unexpected Toxic Deaths ............ 124  
6.6.1.6 Concluding Remarks on Interim Analyses .. 125  
6.7 Phase II/III Trials ......................................... 125  
6.8 Concluding Remark ......................................... 127  

7. Data Management and Quality Control ...................... 129  
7.1 Introduction: Why Worry? ............................... 129  
7.2 Protocol Development ....................................... 133  
7.2.1 Objectives ............................................ 133  
7.2.2 Background ............................................. 133  
7.2.3 Drug Information ..................................... 134  
7.2.4 Stage Definitions .................................... 134  
7.2.5 Eligibility Criteria ................................... 134  
7.2.6 Stratification Factors and Subsets ............... 135  
7.2.7 Treatment Plan ....................................... 135  
7.2.8 Treatment Modification .............................. 135  
7.2.9 Study Calendar ....................................... 136  
7.2.10 Endpoint Definitions ................................. 136  
7.2.11 Statistical Considerations .......................... 137  
7.2.12 Discipline Review ................................... 137  
7.2.13 Registration Instructions ............................ 137  
7.2.14 Data Submission Instructions ....................... 138  
7.2.15 Special Instructions ................................. 138  
7.2.16 Regulatory Requirements ............................ 138  
7.2.17 Bibliography ......................................... 138  
7.2.18 Forms ................................................ 138  
7.2.19 Appendix ............................................. 139  
7.3 Data Collection ............................................ 139  
7.3.1 Basic Data Items ...................................... 140  
7.3.2 Case Report Form Design ............................ 142  
7.4 Data Submission ............................................ 144  
7.4.1 Registration .......................................... 144  
7.4.2 Case Report Forms .................................... 145
## Contents

7.4.3 Specimens ............................................ 146  
7.4.4 Data Submission Enforcement ........................ 146  
7.5 Data Evaluation ........................................... 147  
7.6 Publication ................................................ 150  
7.7 Quality Assurance Audits .................................. 151  
7.8 Training .................................................... 151  
7.9 Database Management .................................... 152  
7.9.1 Database Structures ................................... 152  
7.10 Conclusion ................................................. 153  

8. Reporting of Results .......................................... 155  
8.1 Timing of Report ........................................... 155  
8.1.1 Phase II Trials ......................................... 156  
8.1.2 Phase III Trials ......................................... 157  
8.2 Required Information ...................................... 157  
8.2.1 Objectives and Design ................................ 157  
8.2.2 Eligibility and Treatment .............................. 158  
8.2.3 Results .................................................. 158  
8.3 Analyses .................................................... 159  
8.3.1 Exclusions, Intent to Treat ............................ 159  
8.3.2 Summary Statistics: Estimates and Variability  
of Estimates ................................................... 161  
8.3.3 Interpretation of Results ............................... 164  
8.3.3.1 One-Sided versus Two-Sided Tests ............... 164  
8.3.3.2 Positive, Negative, and Equivocal Trials ........ 165  
8.3.3.3 Multiple Endpoints .................................. 166  
8.3.4 Secondary Analyses .................................... 166  
8.4 Conclusion ................................................. 167  

9. Pitfalls ....................................................... 169  
9.1 Introduction ............................................... 169  
9.2 Historical Controls ....................................... 169  
9.3 Competing Risks ......................................... 177  
9.4 Outcome by Outcome Analyses ........................... 182  
9.4.1 Survival by Response Comparisons .................. 183  
9.4.2 "Dose Intensity" Analyses ............................. 185  
9.5 Subset Analyses ......................................... 190  
9.6 Surrogate Endpoints ..................................... 193  

10. Exploratory Analyses ........................................ 197  
10.1 Introduction ............................................... 197  
10.2 Some Background and Notation .......................... 197  
10.3 Identification of Prognostic Factors .................... 200  
10.3.1 Scale of Measurement ................................. 200  
10.3.2 Choice of Model ....................................... 204
Contents

10.4 Forming Prognostic Groups ........................................ 206
10.5 Analysis of Microarray Data ..................................... 209
10.6 Meta-Analysis ..................................................... 213
  10.6.1 Some Principles of Meta-Analyses ......................... 213
  10.6.2 An Example Meta-Analysis: Portal Vein Infusion ...... 214
    10.6.2.1 Inclusion of Trials ................................... 214
    10.6.2.2 Use of Raw Data ..................................... 215
    10.6.2.3 Lumping Interventions ............................... 215
    10.6.2.4 Quality of Trials .................................... 215
  10.6.3 Conclusions from the Portal Vein Meta-Analysis ....... 216
  10.6.4 Some Final Remarks on Meta-Analysis ................... 217
10.7 Concluding Remarks ............................................. 217

11. Summary and Conclusions ........................................... 219

References ......................................................................... 223

Index .................................................................................. 243
This page intentionally left blank
1

Introduction

It is best to prove things by actual experiment; then you know; whereas if you depend on guessing and supposing and conjectures, you never get educated.

—Mark Twain

... statistics are curious things. They afford one of the few examples in which the use ... of mathematical methods tends to induce a strong emotional reaction in non-mathematical minds. This is because statisticians apply, to problems in which we are interested, a technique which we do not understand. It is exasperating, when we have studied a problem by methods that we have spent laborious years in mastering, to find our conclusions questioned, and perhaps refuted, by someone who could not have made the observations himself.

—Sir Austin Bradford Hill (1937)

1.1 A Brief History of Clinical Trials

The history of clinical trials before 1750 is easily summarized: There were no clinical trials. The basic philosophy of medicine from the time of Hippocrates to the seventeenth century was humoralistic; the accepted version of this philosophy was due to the Greek Galen (130 AD). Since he “laid down ... all that could possibly be said on medicine, he attained an authority that remained unchallenged until well into the sixteenth century. His views on cancer continued to be decisive for an even longer time” (De Moulin, 1989). Illness was caused by imbalances in blood, phlegm, black bile, and yellow bile; treatment consisted of restoring balance. Cancer was caused by a congestion of black bile; appropriate treatment was therefore rigorous purging, a strict bland diet, and, for non-occult disease, poultices and possibly surgery with evacuation of melancholic blood. No matter that the treatments didn’t work—after all, preoccupation with staying alive was contumuously worldly. (Besides, there were always the miracles of Saint Cosmas and Saint Damian if the doctors couldn’t do anything.) Not until the Renaissance were the humoralist bases questioned. Various chemical, mechanical, and electrical causes of cancer were then proposed, and treatments devised in accordance with these causes. Sadly, these treatments were just as ineffective as the theories were inaccurate (e.g.,
arsenic to neutralize acidic cancer juice, diets to dissolve coagulated lymph, bloodletting or shocks to remove excessive electrical irritability). It never occurred to anyone to test whether the treatments worked.

The value of numerical methods began to be appreciated in the 1800s “when in 1806, E. Duvillard in Paris, applying a primitive statistical analysis, showed the favorable effect of smallpox vaccination on the general mortality rate” (De Moulin, 1989, from Duvillard, 1806). These early methods did uncover important epidemiologic facts, but were not so useful in judging treatment effectiveness. Although patient follow-up became the norm rather than the exception, and theories became more sophisticated, typical treatment research consisted only of reports of case series. In an early example of the hazards of such research, reports of the post-operative cure of breast cancer of two Edinburgh surgeons in the 1700s (one the student of the other) were wildly divergent: One was reported as curing 4 out of 60 patients, the other as curing 76 out of 88 (De Moulin, 1989, from Monro, 1781 and Wolff, 1907). Little wonder it was nearly impossible to tell what worked and what did not.

If some of the treatments hadn’t been actively harmful, perhaps it wouldn’t have mattered. Despite major advances in the understanding of disease by 1900, there were still few effective treatments. “The thick textbooks of 1900 are as sweepingly accurate on diagnosis as today’s, but the chapters are all tragedies because they lack a happy ending of effective treatment” (Gordon, 1993). Still, the few medical advances (mercury for syphilis, digitalis for the heart disease, iodine for goitre) and especially the advances in surgery allowed by anesthetics and antiseptics ushered in the “golden age” of Western medicine. Doctors had a priestly role as wise and trusted advisors, with warm and personal relationships with their patients (Silverman, 1992). Of course, these trusted advisors with warm personal relationships didn’t experiment on their patients. Thus even though most of the principles of comparative trials had been enunciated as early as 1866 by Claude Bernard—“... else the doctor walks at random and becomes sport of illusion” (Boissel, 1989, quoted from Bernard, 1866)—and despite the development of modern methods of experimental design in other scientific fields, clinical research remained limited.

In the middle of this century, treatment options began to catch up with biological advances: Questions abounded, and clear answers were not coming fast enough. The first randomized therapeutic clinical trial (1946–48) was the result of a pressing medical problem (tuberculosis), a severely limited supply of a new agent (streptomycin), and frustration with the uninterpretability of 100 years of uncontrolled experimentation. Sir Austin Bradford Hill made the statistical arguments for the trial: The best way to get an answer, particularly given a streptomycin supply sufficient for only 50 patients, was a strictly controlled trial (Hill, 1990). Dr. Phillip D’Arcy Hart, an expert in the treatment of tuberculosis, gave the medical arguments. “The natural course of pulmonary tuberculosis is... so variable and unpredictable that evidence of improvement or cure following the use of a new drug in a few cases cannot be accepted as proof of the effect of that drug. The history of chemotherapeutic trials in tuberculosis is filled with errors...” He went on to note that the
claims made for gold treatment, which had persisted over 15 years, provided a “spectacular example” and concluded that results in the future could not be considered valid unless tested in an adequately controlled trial (Gail, 1996, quoting from Streptomycin in Tuberculosis Trials Committee of the Medical Research Council, 1948).

This first trial demonstrated convincingly that a regimen of streptomycin plus bed rest was superior to bed rest alone. Not at all bad for the first attempt: 15 years and still no answer on gold with the old observational methods, 2 years with the new methods and a clear answer on streptomycin.

The trial of streptomycin for pulmonary tuberculosis “can be seen to have ushered in a new era of medicine,” and Hill generally is agreed to have done “more than any other individual to introduce and foster adoption of the properly randomized controlled trial in modern clinical research” (Silverman and Chalmers, 1991). His efforts to explain and promote good clinical research were tireless and ultimately effective, particularly after the thalidomide tragedy of the 1960s demonstrated the potential harm in not doing carefully controlled trials.

Controlled trials in cancer in the United States were first sponsored by the National Cancer Institute (NCI) under the leadership of Dr. Gordon Zubrod. Zubrod, profoundly influenced by the streptomycin trial, employed the new methods himself (with others) in the study of penicillin in pneumonia and introduced the methods to other early leaders in the cancer clinical trials effort (Zubrod, 1982). Upon his move to the NCI, a comparative study in childhood acute leukemia was designed. This effort expanded into two of the initial cooperative cancer clinical trials groups, the Acute Leukemia Groups A and B; Group B (which became Cancer and Leukemia Group B, or CALGB, recently merged into the Alliance for Clinical Trials in Oncology, or ACTION) had the honor of publishing the first trial (Frei et al., 1958). Zubrod was also instrumental in the formation, in 1955, of the Eastern Solid Tumor Group (now the Eastern Cooperative Oncology Group, or ECOG), which published the first randomized trial in solid tumors in the United States in 1960 (Zubrod et al., 1960).

Of course not everyone was immediately persuaded that randomized trials were the best way to conduct clinical research. Jerome Cornfield, who advised Zubrod, was a major figure in the development of biostatistical methods at the NCI and an early advocate of randomization. His response to the suggestion from a radiotherapist that patients be assigned to conventional therapy or super voltage according to hospital instead of by randomization is often quoted. The quote is a very tactfully worded suggestion that the approach would be suitable if there were no other design options. He ended with an example of a seasickness trial with treatment assigned by boat. How could the trial be interpreted if there were a great deal more turbulence and seasickness on one of the boats? The radiotherapist got the point and randomized by patient (Ederer, 1982). Cornfield is also important for his advocacy of adequate planning, attention to quality, and especially adequate sample size: “... clinical research ... is littered with the wrecks of studies that are inconclusive
or misleading because they were of inadequate scope” (Ederer, 1982, quoting from a memorandum to the Coronary Drug Project steering committee).

Ever since the streptomycin trial, randomized studies have been invaluable in assessing the effectiveness of new therapies. In some cases cherished beliefs have been challenged. An early example was the University Group Diabetes Project (UGDP), which contradicted the widespread view that lowering blood sugar with oral hypoglycemic drugs prolonged life in patients with diabetes. Other examples include the National Surgical Adjuvant Breast and Bowel Program (NSABP) trials of breast cancer surgery demonstrating that more is not better, the Cardiac Arrhythmia Suppression Trial (CAST) trial demonstrating that suppression of ventricular arrhythmia by encaidine or flecainide in patients having recent myocardial infarction increases the death rate instead of decreasing it, and the Southwest Oncology Group (SWOG) trial in non-Hogkin’s lymphoma demonstrating that new highly toxic combination chemotherapy regimens introduced in the 1980s are not better than the old standard combination regimen. However, cherished beliefs die hard. Results such as these met with heavy resistance despite the randomized designs (for other examples see Klimt, 1989); think how easy it would have been to dismiss the results if the designs had been inherently biased. Positive results are happier examples of the importance of clinical trials: The Diabetic Retinopathy Trial demonstrating dramatically reduced visual loss due to photocoagulation therapy, trials establishing the effectiveness of therapy in improving survival in patients with Wilms’ tumor and other childhood cancers, beta blockers prolonging life after myocardial infarction, chemo-radiotherapy substantially improving survival in nasopharyngeal and gastric cancers.

Randomized trials cannot answer every treatment question. Randomization is not feasible in every setting, costs may be prohibitive, and political realities may interfere. Since only a limited number of trials can be done, some questions have to be addressed in other ways. However, controlled trials are by far the best method available for addressing difficult and controversial questions in a way that minimizes distrust of the results. Consider the 1954 Salk Vaccine trial for which at the beginning “the most urgent business was to ... turn the focus away from professional rivalries, power struggles, and theoretical disputes and back to the neglected question of whether or not Salk’s vaccine worked.” Thomas Francis, Jr., was given the job of evaluating the vaccine because “everybody knew that when Tommy Francis talked about working up to a standard, it was one of unimpeachable thoroughness; even the most dedicated opponent to the new vaccine could never say a trial supervised by Francis was political, biased, or incomplete” (Smith, 1990). His two nonnegotiable demands before agreeing to take on the job were that the vaccine proponents would not design the trial and would have no access to the results while the trial was ongoing, and that the trial would have a randomized double blind design instead of an “observed-control” design, in which second graders would have gotten the vaccine and would have been compared to unvaccinated first and third graders (Smith, 1992, quotes from Smith, 1990). The results of this “textbook model of elegant clinical testing”
were unquestionable. Francis’s announcement that “the new vaccine was safe, effective, and potent . . . was a landmark in 20th century history, one of the few events that burned itself into the public consciousness because the news was good” (Smith, 1992). Unimpeachable thoroughness, nonpolitical, unbiased, complete, independent, properly designed—Francis set a very high standard indeed, and one to which all of us involved in clinical research should aspire.

1.2 The Southwest Oncology Group (SWOG)

There are now dozens of national and international consortia of institutions and investigators organized for the purpose of improving the survival of cancer patients through clinical research. Our own experience is with the Southwest Oncology Group, now known simply as SWOG, which began in 1956 in the United States as the (pediatric) Southwest Cancer Chemotherapy Study Group under the direction of Dr. Grant Taylor at the M.D. Anderson Cancer Center in Houston, Texas. In 1958, membership was extended to include investigators evaluating adult malignancies, and in the early 1960s a Solid Tumor Committee was established. Since then the pediatric part of the Group was split off (to become the Pediatric Oncology Group, now part of the Children’s Oncology Group), the name was changed to the Southwest Oncology Group (SWOG), and the Group has expanded to include specialists in all modalities of cancer therapy and institutions in all regions of the country. Most of the studies done by the Group are designed to assess whether a regimen merits further study (Phase II), or to compare two or more regimens (Phase III). Studies in cancer control research (prevention, symptom control and quality of life, survivorship) are also carried out. Currently the Group is led by the Group Chair, Dr. Laurence Baker (University of Michigan, Ann Arbor) and the Group Statistician, Dr. John Crowley (Cancer Research and Biostatistics and Fred Hutchinson Cancer Research Center, Seattle).

The structure of SWOG is typical of cooperative groups and includes the group chair’s office (administration, grants management, industry contracts, legal matters), the operations office (protocol development, meeting planning, regulatory requirements, audits); the statistical center (study development, data base management, network services, computer applications, study quality control, statistical analysis and statistical research); disease committees (breast cancer, gastrointestinal cancer, genitourinary cancer, leukemia, lung cancer, lymphoma, melanoma, myeloma); discipline committees (such as radiotherapy, surgery, nursing and clinical research associates); cancer control and prevention committees (prevention, molecular epidemiology, outcomes and comparative effectiveness, survivorship, and symptom control and quality of life); and all of the participating institutions that enter patients on trials. Group trials and related scientific investigations are proposed and developed within the disease committees under the leadership of the disease
committee chairs. Committee leadership is also provided by the disease committee statistician, who is responsible for reviewing, designing, monitoring and analyzing all studies done in the committee. Each study developed has a clinician assigned (the “study coordinator”) to lead the development effort, to evaluate the data after the study is open, and to be the primary author on the manuscript when the study is complete. Each study also has a “protocol coordinator” assigned from the operations office, who coordinates the production and review of the study protocol, and a “data coordinator” from the statistical center who does most of the necessary setup work for opening a study and reviews and evaluates all of the study data. Participating physicians and clinical research associates at Group institutions are responsible for submitting protocols to their Institutional Review Board, identifying patients suitable for studies, obtaining informed consent, ensuring study participants are treated and followed according to protocol, and for correctly submitting all required data.

The Group typically has 80–100 actively accruing studies at any one time and 400 closed studies in active follow-up. Over 4000 physicians from more than 400 institutions participate. Since the Group began, over 150,000 patients have been registered to its studies, and over 2,000 abstracts and manuscripts have been published. The Group’s extensive clinical trials experience provides the context and many of the examples for this book.

1.3 The Reason for This Book

Our motivations for writing this book are captured by the introductory quotes. Among the four of us we have devoted over 100 years to clinical trials research. As suggested by the first quote, we want to know whether treatments work or not. Furthermore, we want the methods we use to find out whether treatments work to be unimpeachable. Unfortunately, as suggested by the second quote, as statisticians we too often find our motives and methods misunderstood or questioned—and at times actively resented. With this book, it is our hope to improve the mutual understanding by clinicians and statisticians of the principles of cancer clinical trials. Although many of the examples we use are specific to SWOG, the issues and principles discussed are important in cancer clinical trials more generally, and indeed in any clinical trials setting.
2

Statistical Concepts

To understand God’s thoughts we must study statistics, for these are the measure of His purpose.

—Florence Nightingale

2.1 Introduction

A collaborative team that includes both clinicians and statisticians is crucial to the successful conduct of a clinical trial. Although the statistical study design and analyses are mainly the responsibility of the statistician, an understanding of the basic statistical principles is vital for the clinicians involved with the study. The main goal of this chapter is to present statistical concepts that are of particular application to cancer clinical trials.

The objectives of the trial, the key types of data that are collected to meet these objectives, and the types of analyses to be performed are in large part determined by the type of study being undertaken. Phase II trials (discussed in Chapter 5) are small studies early in the development of a regimen that typically focus on toxicity and short-term efficacy data, while Phase III trials (discussed in Chapter 6) are large comparative studies that most frequently assess survival and progression-free survival. We will introduce statistical concepts within the context of each of these types of studies. Phase I trials (discussed in Chapter 4) are a third type of clinical trial which involve a much smaller number of patients, and as such are less suited for use in illustrating basic statistical principles.

First, however, there are some general characteristics of data from clinical trials that do not depend on the type of study. Outcome measures can be classified as being either (1) categorical (qualitative) or (2) measurement (quantitative).

1. Categorical data—outcomes that can be classified according to one of several mutually exclusive categories based on a predetermined set of criteria. For example, standard criteria for solid tumor response (RECIST 1.1, Eisenhauer et al., 2009) categorize patients as achieving
either a CR (complete response) if there is normalization of malignant nodes and disappearance of all other disease; a PR (partial response) if the sum of the diameters of target measurable lesions (shortest diameter for nodes, longest for other lesions) reduces by 30% or more from baseline; INC (increasing) if the diameters increase by 20% or more, or if new sites of tumor are noted; and STA (stable) if none of the above occur. Thus, in this case, patient response can be described by one of four categories, which are then often dichotomized into two categories for analysis (CR + PR versus others; INC versus others).

2. Measured data—outcomes that are measured quantities. For example, concentrations of CA-125, a tumor antigen, are routinely collected in trials of ovarian cancer. Levels of this antigen range in value from 0 to over 10,000. In this case, the data measure a quantity that takes on many possible values. An important special case of quantitative data is time to event data, such as survival time, a measurement in units of time from entry on a study until death. What distinguishes this outcome, and its analysis, from other quantitative data is the frequent presence of what statisticians call censoring. In a typical clinical trial, not all patients have died by the time the study is completed and analyses are performed. For patients still alive, we know that they have lived at least as long as the time from the patient’s entry on the study to the time of the analysis, but we don’t know the actual death time. Special statistical techniques have been developed to incorporate these right censored observations into the analysis.

Three other general concepts we introduce in this section are probability, statistic, and distribution. The probability of an outcome is how often that outcome occurs in relation to all possible outcomes. For instance, the set of all possible outcomes for a flip of a coin is \{H, T\}. The probability of outcome T (tails) is 1/2 if the coin is fair. If the coin is unfair (or biased) and the outcome H (heads) is twice as likely as T, the set of all possible outcomes remains \{H, T\} but the probability of T is now 1/3. If the coin is flipped twice, the set of possible outcomes is \{HH, HT, TH, TT\}. Note that there are two events that yield exactly one tail in two flips, since the tail can be the outcome of either the first or the second flip. With the fair coin the probability of TT is 1/2 × 1/2 = 1/4; with the biased coin it is 1/3 × 1/3 = 1/9. The probability of exactly one tail is (1/2 × 1/2) + (1/2 × 1/2) = 1/2 or (1/3 × 2/3) + (2/3 × 1/3) = 4/9 for the fair and biased coin, respectively. Multiplying the probabilities on each flip to arrive at a probability for the outcomes of events such as TT or HT is justified by an assumption of independence of coin flips, i.e., that the probability of tails on the second flip is unaffected by the outcome of the first flip.

Most commonly used statistical procedures require this independence assumption. What this means is that the value of one observation provides no information about the value of any other. In particular, two measures from the same patient can not be treated as if they were from two separate
(independent) patients. This is because two measurements on the same patient tend to be more alike than two measurements from two different people. Treating multiple observations (such as results from multiple biopsy specimens as to the presence of the multi-drug resistance gene, or MDR) as independent is a common pitfall that should be avoided. For instance, if half of all patients have MDR positive tumors, but within a patient multiple biopsy results are nearly always the same, then about $3/6$ patients would be expected to have MDR positive tumors if all 6 biopsies were from different patients, while either $0/6$ or $6/6$ biopsies would be expected to be positive if all 6 were from the same patient.

The outcome “number of MDR positive tumors out of N tumors” is an example of a statistic, that is, a summary of results from N separate tumors. In general, a statistic is any summary from a set of data points. For example, in the case of measured data such as CA-125, one could summarize the measures from a group of patients using descriptive statistics such as a mean or a median. The statistics chosen to summarize a data set will depend on the type of data collected and the intended use of the information. Some statistics are merely descriptive; others, called test statistics, are used to test hypotheses after the data from an experiment are collected.

A distribution characterizes the probabilities of all possible outcomes of an event or all possible values of a statistic. For a single flip of a fair coin the distribution is

<table>
<thead>
<tr>
<th>outcome</th>
<th>$H$</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>probability</td>
<td>$1/2$</td>
<td>$1/2$</td>
</tr>
</tbody>
</table>

while for the biased coin it is

<table>
<thead>
<tr>
<th>outcome</th>
<th>$H$</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>probability</td>
<td>$2/3$</td>
<td>$1/3$</td>
</tr>
</tbody>
</table>

When a coin is flipped multiple times, a statistic often used to summarize the outcome is the number of tails observed. The distribution of this statistic is the binomial distribution, the most important distribution for categorical data. The distribution from an experiment of flipping a coin $N$ times is characterized by giving the probability that the number of tails is equal to $k$, for every $k$ from 0 to $N$. If the probability of a tail on one flip is $p$, and the flips are independent, the probability of any sequence of flips yielding exactly $k$ tails (and thus $N - k$ heads) is $p^k (1 - p)^{N - k}$. The rest of the exercise is to figure out how many sequences of heads and tails have exactly $k$ tails, and add those up. For $N = 2$ and $k = 1$, we saw above that there were two such sequences, $HT$ and $TH$. In general, the answer is given by the formula $\binom{N}{k} = \frac{N!}{k!(N-k)!}$, where $\binom{N}{k}$ is read “$N$ choose $k$” and $N!$ is $N$ “factorial,” which is simply $N \times (N - 1) \times \cdots \times 2 \times 1$. Thus the probability of exactly three tails in six flips of a fair coin is $\binom{6}{3}(1/2)^3(1 - 1/2)^{6-3} = \frac{6!}{3!3!}(1/2)^6 = 20/64 = 5/16 = 0.3125$. The entire binomial distribution for $N = 6$, and $p = 1/2$ or $1/3$ is given in Figure 2.1 (part (a) for $p = 1/2$; part (b) for $p = 1/3$).
Binomial distributions apply in countless settings other than coin flips; the one of most interest to us in this book is tumor response (the number of responses in N patients taking the place of the number of tails in N flips). In the MDR example above, the binomial distribution applies only to the case where all biopsies are from different patients, since independence is a requirement for the distribution. The probability of exactly three MDR positive patients out of six, if the probability is 1/2 for an individual patient, is 0.3125; if all six biopsies are from the same patient this probability is close to 0. Applying the binomial distribution to the second case is clearly inappropriate.
When outcomes are categorical, the distribution can be shown in a simple table or graph, as above. When outcomes are measured, the distribution can’t be described in a table. Instead, cumulative probabilities are described by a function $F(t)$. For time to death, for example, $F(t)$ means the probability of dying before time $t$. The derivative of $F(t)$, denoted $f(t)$ and often called the density, can be thought of loosely as the probability of dying at $t$ (in a sense made precise by the calculus). We also often talk more optimistically of the survival curve $S(t) = 1 - F(t)$, the probability of surviving at least to time $t$. Note that $S(0) = 1$ and $S(t)$ decreases towards 0 as $t$ gets large. The median survival time, the time past which one half of the patients are expected to live, is that time $m$ for which $S(m) = 0.5$.

Yet another quantity of interest is the hazard function or hazard rate, often denoted $\lambda(t)$. This function can be thought of loosely as the probability of death at time $t$ given that the patient is alive just before time $t$: the instantaneous rate of failure. In terms of the other quantities we have described, the hazard function is given by $\lambda(t) = f(t)/S(t)$. Depending upon the type of disease, the hazard rate as a function of time can take on a variety of forms. For example, in a study involving surgery, a patient’s risk of dying may be highest during the post operative period, then decrease for a period of time. A rising hazard function is characteristic of normal mortality as one ages. In an advanced disease trial, the risk of dying may be relatively constant over the time of follow-up. These three types of hazard functions are given in Figure 2.2a, with the corresponding survival curves in Figure 2.2b.

The constant hazard case with $\lambda(t) = \lambda$ gives rise to what is called the exponential distribution, for which the survival curve is given by

$$S(t) = \exp(-\lambda t),$$

where $\exp$ is the exponential function. Under the assumption of exponential survival, the median survival $m$ is

$$m = -\ln(0.5)/\lambda,$$

where $\ln$ is the natural logarithm. From this relationship it is easy to note that the ratio of hypothesized median survival times for two treatment arms is equal to the inverse of the ratio of the hypothesized hazards. Figure 2.3 shows the hazard function, survival curve, and median survival for an exponential distribution. Although the assumption of a constant hazard rate may not be correct in practice, it can be used to provide reasonable sample size estimates for designing clinical trials (see Chapter 6).

The most common quantitative distribution is the “normal” or “Gaussian” distribution, or “bell-shaped curve.” The standard normal density $f(x)$, presented in Figure 2.4, is a symmetric curve about the mean value 0. The probability of an outcome being less than $x$ is $F(x)$, the area under the density up to $x$. The area under the whole curve has the value 1. The probability that an observation with this standard normal distribution is negative (zero or smaller) is $1/2$, the area under the curve to the left of 0 ($F(0) = 0.5$). The probability
that an observation is greater than 1.645 is 0.05, and the probability that an observation is between $-1.96$ and 1.96 is 0.95. The beauty of the normal distribution is that as sample sizes get large many common statistics that start out as non-normal attain an approximately normal distribution (or a distribution related to the normal, such as the $\chi^2$ discussed in Section 2.3.1). This fact is embodied mathematically in what is known as the Central Limit Theorem. For instance, the probability of $k$ or fewer tails in $N$ coin flips can be found (approximately) from the normal distribution. This is useful in the development of statistical tests, and in the estimation of sample sizes (see Section 2.5).
The remainder of this chapter has been designed to present the key statistical concepts related to cancer clinical trials. For the most part, formulas will only be presented to provide insight into the use of certain statistical tests and procedures; it is far more important to understand why certain statistical techniques are used, rather than how to use them. We will begin with examples that will be used to illustrate the key concepts.