

**HANDBOOK OF STATISTICS
IN
CLINICAL ONCOLOGY
SECOND EDITION**

HANDBOOK OF STATISTICS IN CLINICAL ONCOLOGY

SECOND EDITION

John Crowley, PhD.

*CEO, Cancer Research and Biostatistics (CRAB)
Director, Southwest Oncology Group Statistical Center
Seattle, Washington, USA*

Donna Pauler Ankerst

*Research Scientist
Institute for Medical Informatics, Biometry and Epidemiology
University of Munich
Munich, Germany*



Chapman & Hall/CRC

Taylor & Francis Group

Boca Raton London New York

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

© 2006 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group

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

International Standard Book Number-10: 0-8247-2339-2 (Hardcover)
International Standard Book Number-13: 978-0-8247-2339-2 (Hardcover)

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

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

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

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

Library of Congress Cataloging-in-Publication Data

Catalog record is available from the Library of Congress

informa

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

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

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

To my wife Catherine Abram Crowley, my patient partner of almost 40 years.

John Crowley

To Lorine, in her valiant battle against colon cancer.

Donna Pauler Ankerst

The idea for the first edition of this book came from Graham Garrett of Marcel Dekker Inc., who, tragically, passed away while the book was in progress. Our thanks to those who encouraged this second edition.

All royalties for the editors will go to Cancer Research and Biostatistics, a non-profit corporation whose mission is to help conquer cancer through the application of biostatistical principles and data management methods.

Preface

The second edition expands the first as a compendium of current statistical approaches to problems facing cancer researchers. During the five years since the first edition of the book appeared, there has been an explosion in the technological capabilities of genomic and proteomic research, which is now firmly implanted in cancer research. This has necessitated the addition of a new section devoted to analysis of high throughput data and bioinformatics. Previous sections of the first edition have been revised to reflect the current state of the art. As in the first edition, the intended audience is primarily statisticians working in cancer and more generally in any discipline of medicine. But it is hoped that oncologists too would find the material accessible and benefit from a rudimentary understanding of the fundamental concepts laid forth in each chapter.

The second edition contains six sections:

1. Phase I Trials. Updated recommendations regarding the standard 3-cohort and continual reassessment approaches along with a new chapter on the design and analysis of pharmacokinetic studies are provided.
2. Phase II Trials. Current experience in designs based on toxicity and response, time to events, selection of therapies for phase III, and innovative ways to combine single-arm phase II experiments are discussed.
3. Phase III Trials. A comprehensive treatment is provided of sample size, factorial designs, and noninferiority studies along with analyses of outcomes and economic endpoints.
4. Prognostic Factors and Exploratory Analysis. Essential guidelines and caveats for inferring effects and accounting for covariates for survival analysis are illustrated through a wealth of examples.
5. Bioinformatics and High-Throughput Data. Analyses of genomic microarray, proteomic, single nucleotide polymorphisms, haplotypes, and general features of high-dimensional data are provided, along with a valuable guide to free software.
6. Interpreting Clinical Trials. This section continues to emphasize the lessons and pitfalls on what can and cannot be concluded from single or multiple clinical trials.

Editors

John Crowley is president and CEO of Cancer Research and Biostatistics (CRAB), Seattle, Washington, director of the Southwest Oncology Group Statistical Center, and a faculty member at the Fred Hutchinson Cancer Research Center and the University of Washington. He is the author or coauthor of more than 250 refereed articles, book chapters, and other publications. Dr. Crowley is a fellow of the American Statistical Association and the American Association for the Advancement of Science, and a member of the International Biometrics Society, the Society for Clinical Trials, the American Society for Clinical Oncology, the American Association of Hematology, and the International Association for the Study of Lung Cancer. He received a B.A. degree (1968) from Pomona College, Claremont, California, and M.S. (1970) and Ph.D. (1973) degrees in biomathematics from the University of Washington.

Donna Pauler Ankerst is an assistant professor at the Institute for Medical Informatics, Biometry and Epidemiology, University of Munich, Germany, and an affiliate faculty member at the Fred Hutchinson Cancer Research Center and the University of Washington in Seattle. She is a member of the American Statistical Association, International Biometrics Society, and the International Society for Bayesian Analysis. She received a B.S. degree (1990) from the University of Texas, Austin, and M.S. (1994) and Ph.D. (1996) degrees in statistics from Carnegie Mellon University, Pittsburgh, Pennsylvania.

Contributors

Garnet Anderson

Fred Hutchinson Cancer Research Center
Seattle, USA

John Bryant

University of Pittsburgh
Pittsburgh, Pennsylvania, USA

Iris Burkholder

German Cancer Research Center
Heidelberg, Germany

Mark R. Conaway

University of Virginia Health System
Charlottesville, Virginia, USA

David V. Conti

University of Southern California
Los Angeles, California, USA

John Crowley

Southwest Oncology Group Statistical Center
Cancer Research and Biostatistics
Seattle, Washington, USA

James J. Dignam

University of Chicago
Chicago, Illinois, USA

S. Eberle

Technical University Munich
Munich, Germany

Lutz Edler

German Cancer Research Center
Heidelberg, Germany

Diane L. Fairclough

University of Colorado Health Sciences Center
Aurora, Colorado, USA

James R. Faulkner

Fred Hutchinson Cancer Research Center
Seattle, Washington, USA

Ziding Feng

Fred Hutchinson Cancer Research Center
Seattle, Washington, USA

Dianne M. Finkelstein

Biostatistics Center
Massachusetts General Hospital
Boston, Massachusetts, USA

Robert Gentleman

Fred Hutchinson Cancer Research Center
Seattle, Washington, USA

Ted A. Gooley

Fred Hutchinson Cancer Research Center
University of Washington
Seattle, Washington, USA

Stephanie Green

Pfizer, Inc.
New London, Connecticut, USA

Douglas Grove

Fred Hutchinson Cancer Research Center
Seattle, Washington, USA

S. Hessler

Bristol-Myers Squibb GmbH
Munich, Germany

Norbert Holländer

University of Freiburg
Freiburg, Germany

Li Hsu

Fred Hutchinson Cancer Research Center,
Seattle, Washington, USA

Naoki Ishizuka

Research Institute, International Medical
Center of Japan
Tokyo, Japan

Kenneth J. Kopecky

Fred Hutchinson Cancer Research Center
Seattle, Washington, USA

Michael LeBlanc

Fred Hutchinson Cancer Research Center
Seattle, Washington, USA

Wendy Leisenring

Fred Hutchinson Cancer Research Center
University of Washington
Seattle, Washington, USA

P. Y. Liu

Fred Hutchinson Cancer Research Center
Seattle, Washington, USA

Simon Lunagomez

The University of Texas
M. D. Anderson Cancer Center
Houston, Texas, USA

Paul Marjoram

University of Southern California
Los Angeles, California, USA

Shigeyuki Matsui

Kyoto University School of Public Health
Kyoto, Japan

Carol M. Moinpour

Fred Hutchinson Cancer Research Center
Seattle, Washington, USA

John Molitor

University of Southern California
Los Angeles, California, USA

James Moon

Fred Hutchinson Cancer Research Center
Seattle, Washington, USA

Satoshi Morita

Kyoto University Graduate School of Medicine
Kyoto, Japan

Peter Müller

The University of Texas
M. D. Anderson Cancer Center
Houston, Texas, USA

John O'Quigley

Institute Curie
Paris, France

Joe Pater

Queen's University
Kingston, Ontario, Canada

Donna Pauler Ankerst

Institute for Medical Informatics,
Biometry and Epidemiology
Munich, Germany

Gina R. Petroni

University of Virginia Health System
Charlottesville, Virginia, USA

Scott D. Ramsey

Fred Hutchinson Cancer Research Center
Seattle, Washington, USA

Tim Randolph

University of Washington
Seattle, Washington, USA

Erik Rasmussen

Cancer Research Biostatistics
Seattle, Washington, USA

M. Reck

Krankenhaus Grosshansdorf
Onkologischer Schwerpunkt
Grosshansdorf, Germany

Gary L. Rosner

The University of Texas
M. D. Anderson Cancer Center
Houston, Texas, USA

Willi Sauerbrei

University of Freiburg
Freiburg, Germany

Martin Schumacher

University of Freiburg
Freiburg, Germany

Guido Schwarzer

University of Freiburg
Freiburg, Germany

M. Kriner

Technical University Munich
Munich, Germany

Jonathan J. Shuster

University of Florida
Gainesville, Florida, USA

Yu Shyr

Vanderbilt University School of Medicine
Nashville, Tennessee, USA

Richard J. Stephens

Medical Research Council Clinical Trials Unit
London, UK

Barry E. Storer

Fred Hutchinson Cancer Research Center
Seattle, Washington, USA

Daniel O. Stram

University of Southern California
Los Angeles, California, USA

Catherine M. Tangen

Fred Hutchinson Cancer Research Center
Seattle, Washington, USA

Peter F. Thall

University of Texas
M.D. Anderson Cancer Center
Houston, Texas, USA

Duncan C. Thomas

University of Southern California
Los Angeles, California, USA

Patrick A. Thompson

Baylor College of Medicine
Houston, Texas, USA

Andrea B. Troxel

University of Pennsylvania School of Medicine
Philadelphia, Pennsylvania, USA

K. Ulm

Technical University Munich
Munich, Germany

Xuemei Wang

University of Texas
M.D. Anderson Cancer Center
Houston, Texas, USA

H. Samuel Wieand

University of Pittsburgh
Pittsburgh, Pennsylvania, USA

Ronghui Xu

University of California, San Diego
La Jolla, California, USA

Yutaka Yasui

University of Alberta
Edmonton, Alberta, Canada

Table of Contents

PART I: Phase I Trials

Chapter 1

Overview of Phase I Trials.....	3
<i>Lutz Edler and Iris Burkholder</i>	

Chapter 2

Phase I and Phase I/II Dose Finding Algorithms Using Continual Reassessment Method	31
<i>John O’Quigley</i>	

Chapter 3

Choosing a Phase I Design	59
<i>Barry E. Storer</i>	

Chapter 4

Pharmacokinetics in Clinical Oncology: Statistical Issues.....	77
<i>Gary L. Rosner, Peter Müller, Simon Lunagomez, and Patrick A. Thompson</i>	

Chapter 5

Practical Implementation of the Continual Reassessment Method	97
<i>Naoki Ishizuka and Satoshi Morita</i>	

PART II: Phase II Trials

Chapter 6

Overview of Phase II Clinical Trials.....	119
<i>Stephanie Green</i>	

Chapter 7

Designs Based on Toxicity and Response	131
<i>Gina R. Petroni and Mark R. Conaway</i>	

Chapter 8

Phase II Trials Using Time-to-Event Endpoints	143
<i>Catherine M. Tangen and John Crowley</i>	

Chapter 9

Phase II Selection Designs.....	155
<i>P.Y. Liu, James Moon, and Michael LeBlanc</i>	

PART III: Phase III Trials

Chapter 10

On Use of Covariates in Randomization and Analysis of Clinical Trials 167

Garnet L. Anderson, Michael LeBlanc,

P.Y. Liu and John Crowley

Chapter 11

Factorial Designs with Time to Event Endpoints 181

Stephanie Green

Chapter 12

Noninferiority Trials.....191

Kenneth J. Kopecky and Stephanie Green

Chapter 13

Power and Sample Size for Phase III Clinical Trials of Survival.....207

Jonathan J. Shuster

Chapter 14

Early Stopping of Cancer Clinical Trials 227

James J. Dignam, John Bryant, and H. Samuel Wieand

Chapter 15

Design and Analysis of Quality of Life Data 247

Andrea B. Troxel and Carol M. Moinpour

Chapter 16

Economic Analyses Alongside Cancer Clinical Trials 271

Scott D. Ramsey

PART IV: Exploratory Analysis and Prognostic Factors

Chapter 17

Prognostic Factor Studies.....289

Martin Schumacher, Norbert Holländer, Guido Schwarzer, and Willi Sauerbrei

Chapter 18

Statistical Methods to Identify Predictive Factors 335

Kurt Ulm, Monika Kriner, Sonja Eberle, Martin Reck, and Sybill Hessler

Chapter 19

Explained Variation in Proportional Hazards Regression 347

John O'Quigly and Ronghui Xu

Chapter 20

Constructing Prognostic Groups by Tree-Based

Partitioning and Peeling Methods 365

Michael LeBlanc, Erik Rasmussen, and John Crowley

Chapter 21

Clinical Monitoring Based on Joint Models for Longitudinal

Biomarkers and Event Times 383

Donna Pauler Ankerst and Dianne M. Finkelstein

PART V: High-Throughput Data and Bioinformatics

Chapter 22

Some Practical Considerations for Analysis of Spotted Microarray Data 397

Li Hsu, James R. Faulkner, Douglas Grove, and Donna Pauler Ankerst

Chapter 23

Statistical Applications Using DNA Microarrays for Cancer Diagnosis and

Prognosis 419

Shigeyuki Matsui

Chapter 24

Profiling High-Dimensional Protein Expression Using MALDI-TOF:

Mass Spectrometry for Biomarker Discovery 437

Yutaka Yasui, Tim Randolph, and Ziding Feng

Chapter 25

Statistical Approaches for High Dimensional Data Derived from High Throughput

Assays: A Case Study of Protein Expression Levels in Lung Cancer.....457

Yu Shyr

Chapter 26

Spatial Modeling of Multilocus Data.....471

*David V. Conti, Daniel O. Stram, John Molitor, Paul Marjoram,
and Duncan C. Thomas*

Chapter 27

Software for Genomic Data 489

Robert Gentleman

PART VI: Interpreting Clinical Trials

Chapter 28

Bayesian Sensitivity Analyses of Confounded Treatment Effects.....523

Peter F. Thall and Xuemei Wang

Chapter 29

Interpreting Longitudinal Studies of QOL with Nonignorable Dropout.....541

Diane L. Fairclough

Chapter 30

Estimation of Cumulative Incidence in the

Presence of Competing Risks 557

Ted A. Gooley, Wendy Leisenring, John Crowley, and Barry E. Storer

Chapter 31

Pitfalls in the Design, Conduct and Analysis of Randomized

Clinical Trials 567

Richard J. Stephens

Chapter 32

Dose-Intensity Analysis 581

Joe Pater

Chapter 33

Sequential Randomization 589

Joe Pater and John Crowley

Part I

Phase I Trials

1 Overview of Phase I Trials

Lutz Edler and Iris Burkholder

CONTENTS

1.1	Introduction	4
1.2	Tasks, Assumptions, and Definitions	5
1.2.1	Clinical Issues and Statistical Tasks	5
1.2.2	Assumptions	6
1.2.3	Definitions.....	6
1.2.4	Maximum Tolerated Dose	7
1.2.5	Dose-Toxicity Modeling	9
1.3	Design.....	10
1.3.1	Choice of the Dose Levels: Action Space	10
1.3.1.1	Modified Fibonacci Dose Escalation	11
1.3.1.2	Starting Dose	13
1.3.2	Dose Escalation Schemes	13
1.3.2.1	Traditional Escalation Rule	14
1.3.2.2	Random Walk (RW) Designs.....	16
1.3.2.3	Bayesian Designs	17
1.3.2.4	Continual Reassessment Method (CRM)	18
1.3.2.5	Modifications of the CRM.....	19
1.3.3	Sample Size per Dose Level	20
1.3.4	Validation and Comparison of Phase I Designs	20
1.4	Conduct and Evaluation of the Trial	21
1.4.1	Standard Requirements.....	21
1.4.2	Descriptive Statistical Analysis	21
1.4.3	MTD Estimation	21
1.4.4	Pharmacokinetic Phase I Data Analysis	22
1.4.5	Dose Titration and Toxicity-Response Approach.....	23
1.4.6	Regulations by Good Clinical Practice and International Conference on Harmonization (ICH)	23
1.5	Open Problems	24
	Acknowledgments.....	25
	References	26

LIST OF ABBREVIATIONS

AE	adverse event
ADR	adverse drug reaction
AT	acceptable toxicity
AUC	area under the curve
(b)CRM	(bivariate) continual reassessment method
CTC	Common Toxicity Criteria of the National Cancer Institute
DLT	dose limiting toxicity (often CTC grade 3 and higher)
EORTC	European Organization on Research for Treatment of Cancer
EWOC	escalation with overdose control
GCP	good clinical practice
ICH	International Conference on Harmonization
LD ₁₀	lethal dose 10%
MELD ₁₀	minimal effective dose level for 10% death
MFDE	modified Fibonacci dose escalation
MTD	maximum tolerated dose
NCI	U.S. National Cancer Institute
PGDE	pharmacokinetically guided dose escalation
RW	random walk
SAM	stochastic approximation method
(S)TER	(strict) traditional escalation rule
TDL	toxic dose low
UaD	up-and-down design
UAT	unacceptable toxicity (often CTC grade 4)
WHO	World Health Organization

1.1 INTRODUCTION

The phase I clinical trial constitutes a research methodology for the search and establishment of new and better treatment of human diseases. It is the first of the three phases, I–III, that became a “gold standard” of medical research during the second half of the 20th century.¹ The goal of the phase I trial is to define and to characterize the new treatment in humans to set the basis for later investigations of efficacy. Therefore, the safety and feasibility of the treatment are at the center of interest. A positive risk–benefit judgment should be expected such that possible harm of the treatment is outweighed by possible gains in cure, suppression of the disease and its symptoms, and improved quality of life and survival. The phase I trial should define a standardized treatment schedule to be safely applied to humans that is worthy of further investigation for efficacy.

For non-life-threatening diseases, phase I trials are usually conducted on human volunteers, at least as long as the expected toxicity is mild and can be controlled without harm. In life-threatening diseases such as cancer, AIDS and so on, phase I studies are conducted with patients because of the aggressiveness and possible harmfulness of cytostatic treatments, possible systemic treatment effects, and the high

interest in the new drug's efficacy in those patients directly. After failure of standard treatments or in the absence of a curative treatment for seriously chronically ill patients, the new drug may be the small remaining chance for treatment.

The methodology presented below is restricted to the treatment of cancer patients. The phase I trial is the first instance where patients are treated experimentally with a new drug. Therefore, it has to be conducted strictly under the regulations of the Declaration of Helsinki² to preserve the patient's rights in an extreme experimental situation and to render the study ethically acceptable. Section 1.2 provides an outline of the task including the definition of the maximum tolerated dose (MTD), which is crucial for the design and analysis of a phase I trial. Basic assumptions underlying the conduct of the trial and basic definitions for the statistical task are given. The presentation of phase I designs in Section 1.3 distinguishes between the determination of the dose levels (action space) and the choice of the dose escalation scheme (decision options). This constitutes the core of this chapter. Phase I designs proposed during the past fifteen years are introduced there. The sample size per dose level is discussed separately. Validations of phase I trials rely mostly on simulation studies because the designs cannot be compared competitively in practice. Practical aspects of the conduct of a phase I trial, including the choice of a starting dose, are presented in Section 1.4. Individual dose adjustment and dose titration studies are also addressed. Further, basic pharmacokinetic methods are outlined, and regulatory aspects and guidelines are dealt with. The methodology for phase I trials is far from being at an optimal level at present, and work on the improvement of design used at present as well as on the development of designs addressing new medical hypotheses is ongoing. Therefore, Section 1.5 will address practical needs, problems occurring during the conduct of a phase I trial, and future research topics.

1.2 TASKS, ASSUMPTIONS, AND DEFINITIONS

1.2.1 CLINICAL ISSUES AND STATISTICAL TASKS

Clinical phase I studies in oncology are of pivotal importance for the development of new anticancer drugs and treatment regimens.^{3,4} If a new agent has successfully passed preclinical investigations⁵ and is judged ready for application in patients, then the first application to humans should occur within the framework of a phase I clinical trial.⁶⁻⁹ At this early stage, an efficacious and safe dosing is unknown, and information is available at best from preclinical *in vitro* and *in vivo* studies.¹⁰ Beginning treatment at a low dose very likely to be safe (starting dose), small cohorts of patients are treated at progressively higher doses (dose escalation) until drug-related toxicity reaches a predetermined level (dose limiting toxicity = DLT). The objective is to determine the MTD⁸ of a drug for a specified mode of administration and to characterize the DLT. The goals in phase I trials are according to Von Hoff et al.:¹¹

1. Establishment of an MTD
2. Determination of qualitative and quantitative toxicity and of the toxicity profile
3. Characterization of DLT

4. Identification of antitumor activity
5. Investigation of basic clinical pharmacology
6. Recommendation of a dose for phase II studies

The primary goal is the determination of a maximum safe dose for a specified mode of treatment as the basis for phase II trials. Activity against tumors is examined and assessed, but tumor response is not a primary endpoint.

Inherent in a phase I trial is the ethical issue that anticancer treatment is potentially both harmful and beneficial to a degree that depends on dosage. The dose-dependency is at that stage of research not known for humans.¹² To be on the safe side, the treatment starts at low doses that are probably not high enough to elicit a beneficial effect. Even worse, after having passed the clinical drug development program, the experimental drug may appear as inefficacious or of harm only. The dilemma of probably unknowingly underdosing patients in the early stages of a phase I trial has been of concern and has challenged the search for the best possible methodology for the design and conduct of a phase I trial. The goal is to obtain the most information on toxicity in the shortest possible time with the fewest patients.¹³

Important clinical issues in phase I trials are patient selection, identification of factors that determine toxicity, drug schedules, and the determination and assessment of target toxicity.^{6,11} Important statistical issues are the design parameters (starting dose, dose levels, and dose escalation) and the estimation of the MTD.

1.2.2 ASSUMPTIONS

Most designs for dose finding in phase I trials assume that there exists a monotone dose–toxicity relationship and a monotone dose–(tumor) response relationship.¹⁴ This idealized relationship can be depicted as

biologically inactive dose < biologically active dose < highly toxic dose.

Methods considered below apply to adult cancer patients with a confirmed diagnosis of cancer not amenable to established treatment. Usually excluded are leukemia and tumors in children.⁹ Phase I studies in radiotherapy may require further consideration because of long delayed toxicity. The conduct of a phase I trial requires an excellently equipped oncological center with high quality means for diagnosis and experimental treatment, for detection of toxicity, and for fast and adequate reaction in the case of serious adverse events. Furthermore, easy access to a pharmacological laboratory is needed for timely pharmacokinetic analyses. These requirements indicate the advisability of restricting a phase I trial to one center or a very few centers.

1.2.3 DEFINITIONS

Throughout this article we denote the set of dose levels at which patients are treated by $D = \{x_i, i = 1, \dots\}$ assuming $x_i < x_{i+1}$. The dose unit is usually mg/m² body surface area,¹⁵ but depends on the methods or route of application. It is assumed that patients enter the study one after the other numbered by $j, j = 1, 2, \dots$ and that treatment starts immediately after entry under informed consent. Denote by $x(j)$ the dose

level of patient j . Each patient treated by a specific dose is comprehensively examined for the occurrence of toxicities at a prescribed schedule, and it is determined whether a DLT occurred or not. As such, the toxic response of, say, patient j is described by the dichotomous random variable Y_j where $Y_j = 1$ indicates the occurrence of a DLT and $Y_j = 0$ the nonoccurrence. In order to comply with most published articles we denote the dose-toxicity function by $\psi(x, \mathbf{a})$ with a parameter (vector) $\mathbf{a}$:

$$P(Y = 1 | Dose = x) = \psi(x, \mathbf{a}) \quad (1.1)$$

where $\psi(x, \mathbf{a})$ is assumed as a continuous monotone nondecreasing function of the dose x , defined on the real line $0 \leq x < \infty$, with $\psi(0, \mathbf{a}) \geq 0$ and $\psi(\infty, \mathbf{a}) \leq 1$.

Small cohorts of patients of size n_k are treated on a timely consecutive sequence of doses $x^{[k]} \in D$, where $1 \leq n_k \leq n_{\max}$ and $n_{\max}$ is a theoretical limit on the number of patients treated per dose level (e.g., $n_{\max} = 8$) for $k = 1, 2, \dots$ periods of treatment with dose $x^{[k]}$. Notice that a dose $x_i \in D$ may be visited more than once with some time delay between visits, i.e., $x^{[k]} \neq x^{[h]}$ for $k \neq h$ is not assumed. If the treatment at each of these dose levels $x^{[k]}$, $k = 1, 2, \dots$ lasts a fixed time length Δt (e.g., 2 months), the duration of the phase I trial is then approximately Δt times the number of those cohorts in the trial, independent of the number n_k of patients per cohort at level $x^{[k]}$.

1.2.4 MAXIMUM TOLERATED DOSE

The notion of an MTD is defined unequivocally in terms of the observed toxicity data of the patients treated using the notion of DLT under valid toxicity criteria.⁸ Drug toxicity is considered tolerable if the toxicity is acceptable, manageable, and reversible. Drug safety has been standardized for oncological studies by the establishment of the Common Toxicity Criteria, which are now the Common Terminology Criteria (CTC), of the U.S. National Cancer Institute.¹⁶ This is a large list of adverse events (AEs) subdivided into organ/symptom categories that can be related to the anticancer treatment. Each AE can be categorized into one of six classes

1. CTC grade 0, no AE or normal
2. CTC grade 1, mildly (elevated/reduced)
3. CTC grade 2, moderate
4. CTC grade 3, serious/severe
5. CTC grade 4, very serious or life-threatening
6. CTC grade 5, fatal

The CTC grade 5 is not used in the sequel because death usually is taken as a very serious adverse event and is preceded by a CTC grade 4 toxicity. Of course, a death related to treatment has to be counted as DLT. The extensive list of NCI–CTC criteria has replaced the list of the WHO¹⁷ based on an equivalent 0–4 scale. Investigators planning a phase I trial have to identify in the CTC list a subset of candidate toxicities for dose limitation, and they have to fix the grades for which that toxicity is considered to

be dose limiting such that either treatment has to be stopped or the dose has to be reduced. Usually, a toxicity of grade 3 or 4 is considered as dose limiting. That identified subset of toxicities from the CTC list and the limits of grading define the DLT for the investigational drug. Sometimes the list of DLTs is open such that any AE from the CTC catalog of grade 3 and higher related to treatment is considered as a DLT. Notice the recent change of the Common Toxicity Criteria, version 2.0 (published April 30, 1999), to the Common Terminology Criteria for Adverse Events (CTCAE), version 3.0 (published June 10, 2003), now containing 28 main categories with a total of more than 1000 single AEs. Changes in the CTC grade classification have to be taken into account when toxicities are assessed in studies graded under different CTC versions.

During cancer therapy, patients may show symptoms from the candidate list of DLTs not caused by the treatment but by the cancer disease itself or by concomitant treatment. Therefore, the occurrence of any toxicity is judged by the clinician or study nurse for its relation to the investigational treatment. A commonly used assessment scale is as follows:¹⁸

1. Unclear/no judgment possible
2. Not related to treatment
3. Possibly related to treatment
4. Probably related to treatment
5. Definitively related to treatment

Often, a judgment of possibly or more is considered drug related and called an adverse drug reaction (ADR). Therefore, a more strict definition of the occurrence of DLT is when at least one toxicity of the candidate subset of the CTC criteria of grade 3 or higher has occurred that was judged as at least 'possibly' treatment related. Obviously, this definition carries subjectivity, including the choice of the candidate list of CTCs for DLT, assessment of the grade of toxicity, and assessment of the relation to treatment.¹⁹ Uncertainty of the assessment of toxicity has been investigated.^{20,21} When anticancer treatment is organized in treatment cycles, mostly of length 3–4 weeks, DLT is usually assessed retrospectively before the start of each cycle. In phase I studies, final assessment of the DLT typically is made after a minimum of two cycles. If at least one cycle exhibits at least one DLT, that patient is classified as having reached DLT. An unambiguous definition of the assessment rules of individual DLT is mandatory for the study protocol. For the statistical analysis, each patient should be assessable at his/her dose level either as having experienced a DLT ($Y = 1$) or not ($Y = 0$).

With the above definition of DLT one can theoretically assign to each patient an individual MTD (I-MTD) as the highest dose that can be administered safely to that patient: the I-MTD is the highest dose x that can be given to a patient before a DLT occurs. No within-patient variability is considered in this instance. Because a patient is usually examined at only one dose, it is not possible to observe the I-MTD. The observation is only if the given dose x exceeded the I-MTD or not. It is implicitly assumed that all patients entering a phase I trial react in a statistical sense identically and independently from each other. A population of patients gives rise to a statistical distribution. One postulates the existence of a population-based random MTD

(realized in I-MTDs) to describe this distribution. The probability that x exceeds the random MTD is

$$P(x > \text{MTD}) = F_{\text{MTD}}(x) \quad (1.2)$$

and describes the proportion of the population showing a DLT when treated by dose x . This then becomes the adequate statistical model for describing the MTD. This probabilistic approach, known as a tolerance distribution model for a quantal dose–response relationship,²² allows any reasonable cumulative distribution function F for the right side of Eq. (1.2). F is a nondecreasing function with values between 0 and 1. In practice one should allow $F(0) > 0$ as a “baseline toxicity” and also $F(x_{\max}) < 1$ for saturation of toxicity. Classes of well-known tolerance distributions are the probit, logit, and Weibull models.²² On this probabilistic basis, a practicable definition of an MTD of a phase I trial is obtained as a percentile of the statistical distribution of the (random) MTD as follows.

Determine an acceptable proportion $0 < \theta < 1$ of tolerable toxicity in the patient population before accepting the new anticancer treatment. Define the MTD as that dose x_{MTD} for which the proportion of patients exceeding the DLT is at least as large as θ : $F_{\text{MTD}}(x_{\text{MTD}}) = \theta$ or $x_{\text{MTD}} = F_{\text{MTD}}^{-1}(\theta)$. Obviously, there is a direct correspondence between F in Eq. (1.2) and ψ in Eq. (1.1) as

$$\psi(x_{\text{MTD}}, \mathbf{a}) = P(Y = 1 | \text{Dose} = x_{\text{MTD}}) = F_{\text{MTD}}(x_{\text{MTD}}) = \theta \quad (1.3)$$

If $\psi(x, \mathbf{a})$ is monotone nondecreasing in x and continuous, the MTD for θ , denoted MTD_{θ} , is the θ -percentile:

$$\text{MTD}_{\theta} = \psi^{-1}(\theta, \mathbf{a}) \quad (1.4)$$

Figure 1.1 shows how the MTD_{θ} is calculated. The choice of θ depends on the nature of the DLT and the type of the target tumor. For an aggressive tumor and a transient and non-life-threatening DLT, θ could be as high as 0.5. For persistent DLT and less aggressive tumors, it could be as low as 0.1 to 0.25. A commonly used value is $\theta = 1/3 = 0.33$.

1.2.5 DOSE-TOXICITY MODELING

The choice of an appropriate dose-toxicity model $\psi(x, \mathbf{a})$ is important not only for the planning but also for the analysis of phase I data. Most applications use an extended logit model because of its flexibility; the ease of accounting for patient covariates, such as pretreatment, disease staging, and performance; and the availability of computing software. A general class of dose-toxicity models is given by the two-parameter family

$$\psi(x, \mathbf{a}) = F(a_0 + a_1 h(x)) \quad (1.5)$$

where F is a known cumulative distribution function, h a known dose metric, and $\mathbf{a} = (a_0, a_1)$ two unknown parameters to be estimated. Monotone increasing functions F and h are sufficient for a monotone increasing dose-toxicity function ψ .

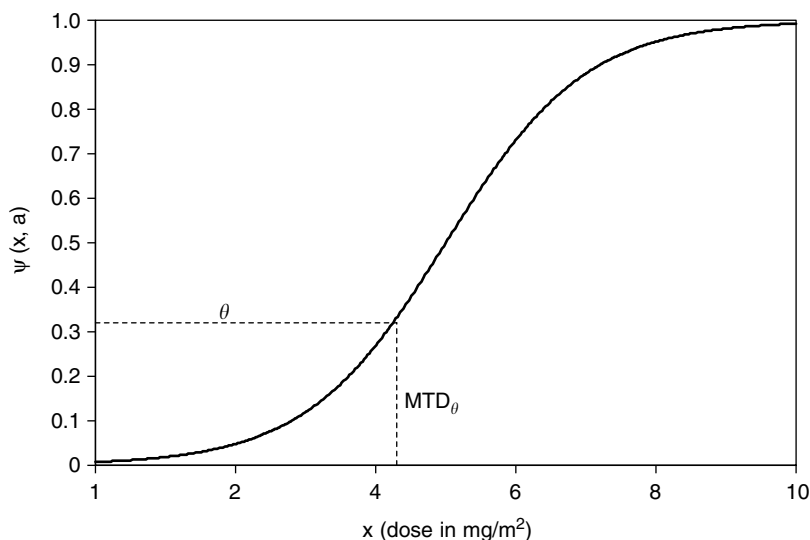


FIGURE 1.1 Schematic dose-toxicity relationship $\psi(x, 5) = \exp(-5 + x)/(1 + \exp(-5 + x))$. The maximum tolerable dose MTD_θ is defined as the θ percentile of the monotone increasing function $\psi(x, a)$ with model parameter $a = 5$.

If $h(x) = x$, the MTD is

$$MTD_\theta = \frac{F^{-1}(\theta) - a_0}{a_1} \quad (1.6)$$

Convenient functions F are the

Probit (x): $\Phi(x)$

Logit (x): $\{1 + \exp(-x)\}^{-1}$

Hyperbolic tangent (x): $\{[\tanh(x) + 1]/2\}^{-a_2}$

with a further unknown parameter component a_2 (see O'Quigley et al).²³

1.3 DESIGN

A phase I trial design has to determine which dose levels are applied to how many patients and in which sequel given the underlying goal of estimating the MTD. This implies three tasks: determining of the possible set of dose levels, choosing the dose levels sequentially, and determining the number of patients per dose level.

1.3.1 CHOICE OF THE DOSE LEVELS: ACTION SPACE

From previous information, mostly preclinical results, a range D of possible doses (action space) is assumed. One may distinguish between a continuous set of doses

D_C , a discrete finite action space $D_K = \{x_1 < x_2 < \dots < x_K\}$ of an increasing sequence of doses, or an infinite ordered set $D_\infty = \{x_1 < x_2 < \dots\}$. Simple dose sets are the additive set

$$x_i = x_1 + (i-1) \Delta x \quad i = 1, 2, \dots \quad (1.7)$$

with dose increment Δx and the multiplicative set

$$x_i = x_1 \cdot f^{i-1} \quad i = 1, 2, \dots \quad (1.8)$$

where f denotes the factor by which the starting dose x_1 is increased. A pure multiplicative set cannot be recommended and is not used in phase I trials because of its extreme danger of jumping from a nontoxic level directly to a highly toxic level. Modifications of the multiplicative scheme that start with a few large steps and slow down later are in use. Such a modified action space could be the result of a mixture of Eq. (1.7) and Eq. (1.8), where the first steps at low doses are obtained multiplicatively and the remaining steps additively. Another, smoother set of doses is obtained when the factors are decreasing with higher doses, for example:

$$x_i = f_{i-1} x_{i-1} \quad i = 1, 2, 3, \dots \quad (1.9)$$

where $\{f_i\}$ is a nonincreasing sequence of factors that may start with $f_1 = 2$ as doubling dose from x_1 to x_2 and continues with $1 < f_i < 2$ for $i \geq 2$. The modified Fibonacci scheme described next is of this type. It has been in use from the beginning of systematic phase I research (see Edler).²⁴

1.3.1.1 Modified Fibonacci Dose Escalation

The most popular and most cited dose escalation scheme is the so-called modified Fibonacci dose escalation (MFDE). Table 1.1 shows Fibonacci sequences and modifications that form the basis of this scheme. A review of the literature for its origin and justification as a dose-finding procedure is difficult. A number of authors^{8,25} refer to an article from 1975 of Goldsmith et al.,²⁶ who present the MFDE as an “idealized modified Fibonacci search scheme” in multiples of the starting dose and as percent of increase. Two years earlier Carter³ summarized the study design principles for early clinical trials. For methodology he referred in a general way to O. Selawry, chief of the Medical Oncology Branch at the NCI in the early seventies, “who has elucidated many of the phase I study principles”. Carter stated that “this scheme has been used successfully in two phase I studies performed by the Medical Oncology Branch” in 1970, one by Hansen²⁷ and one by Muggia.²⁸ Both studies are published in the *Proceedings of the American Association of Cancer Research* without a bibliography. However, the relation to the Fibonacci numbers is not clarified in these early papers. In 1977 Carter²⁹ referred to “a dose escalation based on a numeral series described by the famed 13th Century Italian mathematician Leonardo Pisano, alias Fibonacci.” He also reported on the use of the MFDE by Hansen et al.,³⁰ when studying the antitumor effect of 1.3-bis(2-chloroethyl)-1-nitrosourea (BCNU) chemotherapy. Both Carter and Hansen et al. refer to Schneiderman.¹⁰

The weakness of the foundation of the MFDE for phase I trials becomes even more evident when one looks deeper into history.

Fibonacci (Figlio Bonaccio, the son of Bonaccio), also known as Leonardo da Pisa or Leonardo Pisanus, lived from 1180 to 1250 as a mathematician at the court of Friedrich II in Sicily working in number theory and biological applications. The sequence of numbers named after Fibonacci is created by a simple recursive additive rule: each number is the sum of its two predecessors: $f_{n+1} = f_n + f_{n-1}$ and it starts from $f_1 = 1$ using $f_0 = 0$ (Table 1.1). Fibonacci related this sequence to the “breeding of rabbits” problem in 1202 and also to the distribution of leaves about a stem. The sequence f_n grows geometrically and is approximately equal to $a[(1 + \sqrt{5})/2]^n$, where $a = (1 + \sqrt{5})/\sqrt{5}$. The ratio of successive numbers f_n/f_{n-1} converges to $(1 + \sqrt{5})/2 = (1 + 2.236)/2 = 1.618$, the golden section, a famous principle of ancient and Renaissance architecture. The Fibonacci numbers have been used in optimization and dynamic programming in the 1950s³¹ for determining the maximum of a unimodal function. One application can be illustrated as follows: “How many meters long can a bent bridge be, such that one can always locate its maximum height in units of meters by measuring at most n times?” The solution is given by Bellman’s theorem³¹ as f_n meters. The results say nothing on the placement of the measurements but only on the needed number of measurements.

In his article on methods for early clinical trials research, Schneiderman¹⁰ showed that he was familiar with this work on optimization and cites Bellman’s result but gives the wrong page number (correct is (31) page 34, not page 342). The optimization result is even better explained in³² on page 152. Schneiderman¹⁰ tried to transpose the Fibonacci search by fixing an initial dose x_1 , a maximum possible dose x_K , and the number n of steps for moving upwards from x_1 to x_K . “By taking a Fibonacci series of length $n + 1$, inverting the order, and spacing the doses in proportion to the n intervals in the series”,¹⁰ Schneiderman obtained an increasing sequence of doses. In contrast to the MFDE, which is based on a multiplicative set of doses, this approach is somehow still additive but leads to smaller and smaller steps toward higher doses similar to the MFDE. However, the steps obtained by

TABLE 1.1
Fibonacci Numbers and Consecutive Ratios of Fibonacci Numbers as well as Modified Fibonacci Dose Escalation Schemes Used in Clinical Practice²⁴

Fibonacci Numbers $f_n, n = 1, 2, \dots$	Fibonacci Multiples f_{n+1}/f_n	Modified Fibonacci	Smoothed Modified Fibonacci
1	—	—	—
2	2.0	2	2.0
3	1.5	1.65	1.67
5	1.67	1.52	1.50
8	1.60	1.40	1.40
13	1.63	1.33	1.30 – 1.35
21	1.62	1.33	1.30 – 1.35
34	1.62	1.33	1.30 – 1.35
55	1.62	1.33	1.30 – 1.35

Schneiderman's inversion are at the beginning very large and later very small. This escalation is fixed to K doses and does not open to higher doses if the MTD is not reached at x_K . Schneiderman discussed a restarting of the scheme if no toxicity is seen at x_K and concludes that then "no guide seems to exist for the number of steps." And he confesses that he has "not seen it in any published account of preliminary dose finding." The number of steps in this reversed Fibonacci scheme is strongly related to the escalation factor and so provides no guidance for dose escalation. In the same paper, however, he cites a paper of de Vita et al.³³ in which a dose escalation with the factors of 2, 1.5, 1.33, and 1.25 is used. A hint to the use of the inverse of a Fibonacci scheme where the dose increments decrease with increasing numbers is also given by Bodey and Legha,⁸ who refer to²⁶ "who examined the usefulness of the modified Fibonacci method as a guide to reaching the MTD."

In summary, it seems that the idea of the so-called MFDE came up in the 1960s in the NCI where the early clinical trials programs started and that it was promoted by the scientists mentioned above. They searched for a dose escalation scheme that slows from doubling the dose to smaller increases within a few steps. The MFDE (Table 1.1), slowing down the increase from 65% to 33% within the first 5 steps, seemed reasonable enough to be used in many trials. The method has been successful to the extent that MTDs have been determined through its use. From empirical evidence and the simulation studies performed later, the MFDE seems now to be too conservative in too many cases.

1.3.1.2 Starting Dose

The initial dose given to the first patients in a phase I study should be low enough to avoid severe toxicity but also high enough for a chance of activity and potential efficacy. Extrapolation from preclinical animal data focused on the dose with 10% drug-induced deaths in mice (LD_{10}) converted into equivalents in units of mg/m² of body surface area.³⁴ The standard starting dose became 1/10 of the minimal effective dose level for 10% deaths ($MELD_{10}$) in mice after verification that no lethal or life-threatening effects were seen in other species, for example, rats or dogs.^{7,11,35} Earlier recommendations had used higher portions of the $MELD_{10}$ (mouse) or other characteristic doses, such as the lowest dose with toxicity in mammals.³⁶

1.3.2 DOSE ESCALATION SCHEMES

If a clinical action space has been defined as a set of dose levels D , the next step in designing a phase I trial is establishment of the rule by which the doses of D are assigned to patients. Proceeding from a starting dose x_1 , the sequence of dosing has to be fixed in advance in a so-called dose escalation rule. This section starts with the traditional escalation rules (TER). These rules are also known as "3 + 3" rules because it became usual to enter three patients at a new dose level and a maximum of three more when any toxicity was observed¹¹ before deciding to stop at that level or to increase the dose. Carter³ is the first source these authors are aware of to refer the "3 + 3 rule" as a phase I study principle. Two versions of this "3 + 3 rule" are described below as TER and strict TER (STER), respectively. Then we introduce the up-and-down rules (UaD) as fundamental but not directly applicable rules and turn

from them to Bayesian rules and the continual reassessment method (CRM). Methods for the determination of the MTD have to be addressed for each of the rules.

1.3.2.1 Traditional Escalation Rule

A long used standard phase I design has been the TER, where the doses escalate in D_K or D_∞ step by step from x_i to x_{i+1} , $i = 1, 2, \dots$, with three to six patients per dose level. Table 1.2 shows an example in advanced malignancies.³⁷

An example of TER using cohorts of three is shown in Figure 1.2. Using TER, patients are treated in cohorts of three each receiving the same dose, say, x_i . If none of the three patients shows a DLT at level x_i , the next cohort of three patients receives the next higher dose x_{i+1} . Otherwise, a second cohort of three is treated at the same level x_i again. If exactly one out of the six patients treated at x_i exhibits DLT, the trial continues at the next higher level x_{i+1} . If two or more patients out of the six exhibit a DLT at the level x_i , the escalation stops at that level.

When the escalation has stopped, various alternatives of treating a few more patients are in use:

- 1. Treat a small number of additional patients at the stopping level x_i , for example, to a total of eight patients.
- 2. Treat another cohort of three patients at the next lower level x_{i-1} if six patients have not already been treated.
- 3. Treat cohorts of three patients at all next lower levels $x_{i-1}, x_{i-2}, \dots$ if only three patients have been treated earlier, possibly going down as far as x_1 .
- 4. Treat a limited number of patients at a level not previously included in D located between x_{i-1} and x_i .

TABLE 1.2
Example of a Phase I Study Performed According to TER^a

Dose Level	Dosage	Escalation Factor	N	DLT	Grade
1	0.45	—	3	0	000
2	0.9	2	6	1	111311
3	1.5	1.67	6	1	041111
4	3	2	3	0	012
5	4.5	1.5	4	0	1122
6	6.0	1.33	3	0	111
7	7.5	1.25	3	1	113
8	10.5	1.4	3	0	111
9	13.5	1.29	4	1	0320
10	17.5	1.29	3	0	010
11	23.1	1.33	6	1	123122
12	30.0	1.3	5	4	33331
13	39.0	1.3	1	1	3

^aColumns show dose level, dosage (units), escalation factor, number of patients N, number of cases with dose limiting Toxicity (DLT) defined as grade 3–4, and the actually observed toxicity grades of the N patients.³⁷

A slightly more conservative escalation is implemented in a modified TER, denoted as STER and illustrated in Figure 1.3. Using STER, patients are treated in cohorts of three, each receiving the same dose, say x_i . If none of the three patients shows a DLT at level x_i , the next cohort of three patients receives the next highest dose x_{i+1} . If one DLT is observed, three other patients are included at the same level x_i and the procedure continues as TER. If two or three DLTs are observed among the first three of that cohort, escalation stops at x_i and the dose is deescalated to the next lower level x_{i-1} where a predetermined small number of cases is treated additionally according to one of the options.²⁻⁴ Another dose escalation rule, the “best-of-5” design is described by Storer in this volume.³⁸

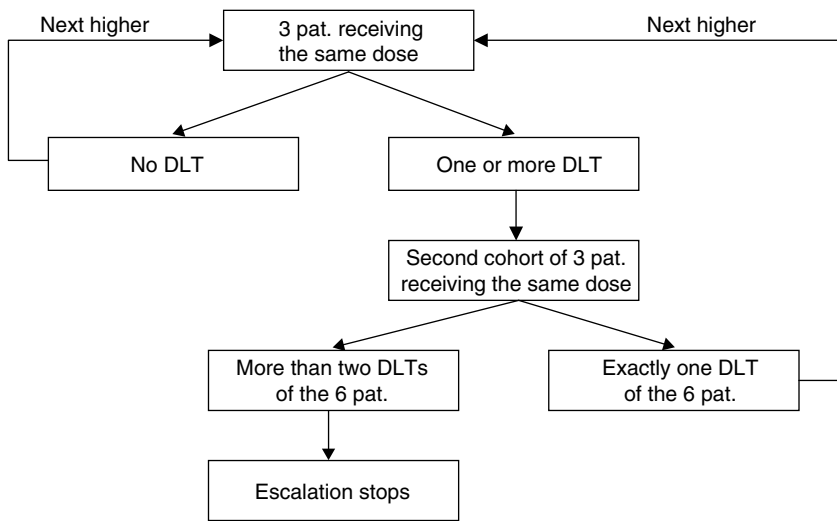


FIGURE 1.2 Traditional escalation rule (TER).

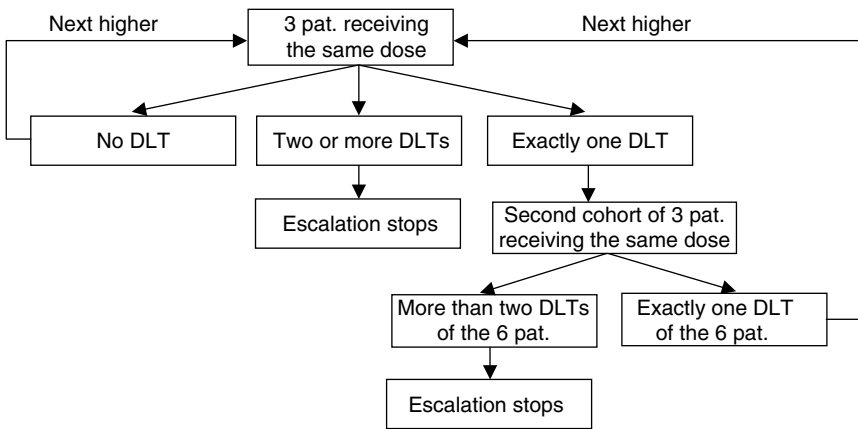


FIGURE 1.3 Strict traditional escalation rule (STER).

If the escalation stops at a dose level x_i , that level would be the first level of unacceptable DLT, and one would conclude that the MTD is exceeded. The next lower level x_{i-1} is then considered as the MTD. It is common practice that at the dose level of the MTD at least six patients are treated. For this reason, options 2 and 3 above are often applied at the end of a phase I trial. Therefore, the MTD can be characterized practically as the highest dose level below the stopping dose level x_i at which at least six patients have been treated with no more than one case of DLT. If no such dose level can be identified, the starting dose level x_1 would be taken as, MTD.

1.3.2.2 Random Walk (RW) Designs

A large class of escalation rules is based on the sequential assignment of doses to one patient after the other. Those rules have their origin in sequential statistical designs and in stochastic approximation theory. If the dose assignment to the current patient depends only on the result seen in the previous patient, the assignment process becomes Markovian and performs a random walk on the action space D . Basically, a patient is assigned to the next higher, the same, or the next lower dose level with a probability that depends on the previous subject's response. RW designs operate mostly on the finite lattice of increasingly ordered dosages $D_K = \{x_1 < \dots < x_K\}$. A Markov chain representation of the random walk on D was at first described for phase I studies by Storer³⁹ (see also).³⁸ In principle, the MTD is estimated after each patient's treatment, with the next patient treated at that estimated level. All optimality results of RW designs require that the set D of doses remains unchanged during the trial. These designs are simple to implement, essentially nonparametric and of known finite and asymptotic distribution behavior. RW designs have been applied to phase I studies.^{39,40} Early prototypes in statistical theory were the up-and-down designs (UaD),⁴¹ proposed originally for explosives testing, and the stochastic approximation method (SAM).⁴² SAM has never been considered seriously for phase I trials. One reason may be the use of a continuum of dose levels leading to impracticable differentiation between doses; another reason could be the ambiguity of the adapting parameter sequence (See Ref. 43). The main reason has been stated already in:¹⁰ "the up and down procedures and the usual overshooting are not ethically acceptable in an experiment on man." However, the UaD rules were more recently adapted for medical applications by considering grouped entry, biased coin randomization, and Bayesian methods.⁴⁰ As such, the elementary UaD has been reintroduced into phase I trials as a tool to construct more appropriate combination designs; see Storer's B-design³⁹ for an example.

Figure 1.4 shows the elementary UaD. Given patient j has been treated on dose level $x(j) = x_i$, the next patient $j + 1$ is treated at the next lower level x_{i-1} if a DLT was observed in patient j , otherwise at the next higher level x_{i+1} .

Two modifications of the elementary rule were proposed:³⁹ modified by two UaD, or Storer's C design (UaD-C), and modified by three UaD, or Storer's D design (UaD-D), which is quite similar to the "3 + 3 rule":

UaD-C: Proceed as in UaD but escalate only if two consecutive patients are without DLT.

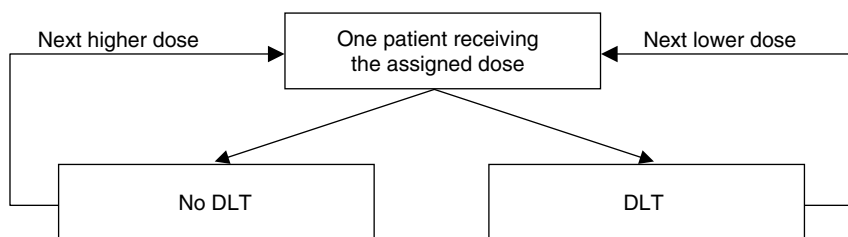


FIGURE 1.4 Elementary up-and-down design (UaD).

UaD-D: Three patients are treated at a new dose level. Escalate if no DLT and deescalate if more than one DLT occurs. If exactly one patient shows a DLT, another three patients are treated at the same level and the rule is repeated.

The two designs were combined with the elementary UaD to Storer's BC and BD two-stage designs:

UaD-BC: Use the UaD until the first toxicity occurs and continue with the UaD-C at the next lower dose level.

UaD-BD: Use the UaD until the first toxicity occurs and continue with UaD-D design. For a detailed description of this two-stage design see.³⁸

Simulations revealed a superiority of the UaD-BD over the UaD-BC and the elementary UaD.³⁹ Although the single-stage designs UaD, UaD-B, and UaD-C are not considered as sufficient and only the two-stage combinations are proposed for use, unfortunately, proposals of new designs have been calibrated mostly on the one-stage designs instead of on the more successful two-stage designs. Therefore, recommendations for practice are hard to deduct from those investigations.

A new sequential RW is the so-called biased coin design (BCD)⁴⁰ applicable for an action space $D_K = \{x_1 < \dots < x_K\}$. The BCD is shown in Figure 1.5. Given patient j has been treated on dose level $x(j) = x_i$, the next patient $j + 1$ is treated at the next lower level x_{i-1} if a DLT was observed in patient j , otherwise at x_i with some probability $p\#$ not larger than 0.5 or at x_{i+1} with probability $1 - p\#$ (not smaller than 0.5). When reaching boundaries x_1 and x_K , the procedure must stay there. This design centers the dose allocation unimodally around the MTD_θ for any θ for $0 < \theta \leq 0.5$ if $p\#$ is chosen as $\theta/(1-\theta)$ and if the MTD is within the action space D_K . Interestingly, the 33rd percentile $MTD_{1/3}$ is obtained with a nonbiased coin of probability $p\# = (1/3)/(2/3) = 0.5$.

1.3.2.3 Bayesian Designs

Bayesian methods are attractive for phase I designs because they can be applied even if few data are present but prior information is available. Furthermore, Bayesian methods are excellently adapted to decision making. They tend to allocate patients to higher doses after nontoxic responses and to lower doses after toxic responses and

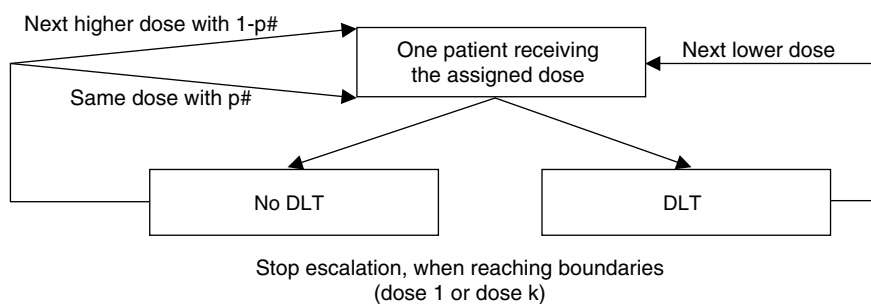


FIGURE 1.5 Biased coin design (BCD).

thus enter into the particulars of the ethical dilemma of phase I trials. Bayesian designs are best suited when patients are treated one at a time and when toxicity information is available quickly. A full Bayesian design⁴⁴ is set up by Eq. (1.1), the prior distribution $g(\mathbf{a})$ of the model parameter $\mathbf{a}$, and a continuous set actions D . A gain function (negative loss) G is needed to characterize the gain of information if an action is taken given the true (unknown) value of $\mathbf{a}$, or equivalently the MTD.

The sequentially collected dose–toxicity information up to the $j - 1$ th patient is summarized by $\Omega_j = \{y_1, \dots, y_{j-1}\}$. The information upon the parameter $\mathbf{a}$ is described by the posterior density distribution of the parameter $\mathbf{a}$ given Ω_j . Using the dose toxicity function (1.1), the likelihood of the data Ω_j for the first $j - 1$ patients is given by

$$f_{j-1}(\Omega_j, \mathbf{a}) = \prod_{s=1}^{j-1} \psi(x(s), \mathbf{a})^{y_s} [1 - \psi(x(s), \mathbf{a})]^{1-y_s} \quad (1.10)$$

Then the posterior distribution is

$$g(\mathbf{a}|\Omega_j) = \frac{f_{j-1}(\Omega_j, \mathbf{a})g(\mathbf{a})}{\int f_{j-1}(\Omega_j, \mathbf{a})g(\mathbf{a})d\mathbf{a}}. \quad (1.11)$$

The next dose $x(j)$ is determined such that the posterior expected gain $E[g(\mathbf{a}) | \Omega_j]$ is maximized with respect to $x(j)$. For details see.⁴⁴ Gatsonis and Greenhouse⁴⁵ used a Bayesian method to estimate directly the dose–toxicity function and the MTD_θ in place of the parameter $\mathbf{a}$. The probability $\alpha = P(Dose > MTD_\theta)$ of overdosing a patient was used as the target parameter for Bayesian estimation in the so-called escalation with overdose control (EWOC) method.⁴³ Whitehead and Brunier⁴⁴ use the precision of the MTD estimate obtained from the next patient as the gain function.

1.3.2.4 Continual Reassessment Method

In order to reduce the number of patients treated with possibly ineffective doses by the TER/STER or the RW type designs a Bayesian-based dose escalation rule was introduced by O’Quigley and co-workers²³ and recommended in a series of follow-up articles.⁴⁶ A starting dose is selected using a prior distribution of the MTD, which is

updated with each patient's observation with regard to the absence or presence of DLT. Each patient is treated at the dose level closest to the currently estimated MTD. Assume a finite action space $D_K = \{x_1 < \dots < x_K\}$ of dose levels, a fixed sample size N , and a one-parameter dose-toxicity function $\psi(x, \mathbf{a})$ as defined in Eq. (1.1). Estimation of the MTD is therefore equivalent to the estimation of $\mathbf{a}$. Assume that there exists a unique solution a^0 at the MTD_θ satisfying $\psi(MTD, a^0) = \theta$. Using the notation of the previous section, assume $f_{j-1}(\Omega_j, \mathbf{a})$ is normalized by its integral over $\mathbf{a}$.

Using CRM, the next dose level is determined such that it is closest to the current estimate of an MTD_θ . For this, the probability of a toxic response is calculated for each $x_i \in D_K$ as

$$P(x_i, \mathbf{a}) = \int_0^{\infty} \psi(x_i, \mathbf{a}) f(\mathbf{a}, \Omega_j) d\mathbf{a} \quad (1.12)$$

and the dose level $x = x(j)$ for the j th patient is selected that minimizes the distance from $P(x, \mathbf{a})$ to the target toxicity rate θ . After observing toxicity Y_j at dose level $x(j)$, the posterior density of the parameter $\mathbf{a}$ is updated by the likelihood of the j th observation. After the planned N patients have been treated, the MTD is estimated as $MTD = x(N+1)$, the dose for the $(N+1)$ st patient. Consistency in the sense that the recommended dose converges to the target level has been shown even under model misspecification.⁴⁷

1.3.2.5 Modifications of the CRM

CRM has been criticized from three aspects: choice of the starting dose $x(1)$ according to a prior $g(\mathbf{a})$ that could result in a dose level in the middle dose region and not in the lower dose region, allowance to jump over a larger part of the dose region, and lengthening of the duration of the trial due to examination of only one patient at a time. Modifications of the CRM to answer these criticisms have been proposed.^{48–52} From these proposals evolved a design sketched in⁵³ relying on a suggestion of Faries.⁴⁹

Using modified CRM, start with one patient at dose level x_1 and apply the CRM. Given patient $j-1$ has been treated on dose level $x(j-1) = x_i$ and information Ω_j predicts for the next patient j the dose level $x_{CRM}(j)$, then the next dose level $x(j)$ is chosen as

$$x(j) = \begin{array}{ll} x_{CRM}(j) & \text{if } x_{CRM}(j) \leq x(j-1) \\ x_{i+1} & \text{if } x_{CRM}(j) > x(j-1) \text{ and if } y_{j-1} = 0 \text{ (no DLT)} \\ x_i & \text{if } x_{CRM}(j) > x(j-1) \text{ and if } y_{j-1} = 1 \text{ (DLT)} \end{array}$$

Further modifications are as follows.

1. Modified CRM but stay with $x(j)$ always one dose level below $x_{CRM}(j)$ (version 1 in)⁴⁹
2. Modified CRM but $x(j)$ is not allowed to exceed the MTD estimate based on Ω_j (version 2 in)⁴⁹

3. Modified CRM but use the starting level of the CRM and enter there three patients (version 4 in)⁴⁹
4. CRM with escalation restricted within $D_K = \{x_1 < \dots < x_K\}$ to one step only (restricted version in)⁵⁰
5. Modified CRM but stopping if the next level has been already visited by a predetermined number of patients. Korn et al.⁴⁸ introduce additionally a dose level $x_0 < x_1$ for an estimate of the MTD if a formally proposed MTD equal to x_1 would exhibit an unacceptable toxicity.
6. Modified CRM run in three variants of one, two, or three patients per cohort.⁵² This version is identical to modification 5 except that a dose level cannot be passed after a DLT was observed.
7. CRM but allow more than one patient at one step at a dose level and limit escalation to one step.⁵¹ A simulation study showed a reduction of trial duration (50–70%) and reduction of toxicity events (20–35%) compared to the CRM.

For additional modifications and details see (See Ref. 46,50,52,54,55,56).

1.3.3 SAMPLE SIZE PER DOSE LEVEL

The number of patients to be treated per dose level was often implicitly determined in the previous described designs. Statistical methods on optimal sample sizes per dose levels seem to be missing. Recommendations vary between one and eight patients per dose level; other suggestions were three to six patients per lower dose levels⁹ or a minimum of three per dose level and a minimum of five near the MTD.⁷ Calculations of sample sizes separately from the sequential design can be based on the binomial distribution and hypothesis testing of toxicity rates.⁵⁷ This gives some quantitative aid in planning the sample size per selected dose level. Using the two probabilities P_{AT} and P_{UAT} of acceptable toxicity (AT) and unacceptable toxicity (UAT), tables for a fixed low $P_{AT} = 0.05$ and a few P_{UAT} values were given by Rademaker⁵⁷ with a straightforward algorithm. Characteristically, the sample size and the error rates increase rapidly if the two probabilities approach each other. If $P_{AT} = 0.05$, a sample size of $n \leq 10$ per dose level is only achieved if $P_{UAT} \geq 0.2$. If $n = 6$, the probability of escalating becomes 0.91 if the toxicity rate equals $P_{AT} = 0.05$, and the probability of nonescalating becomes 0.94 if the toxicity rate equals $P_{UAT} = 0.4$. The latter probability decreases from 0.94 to 0.90, 0.83, and 0.69 if n decreases from 6 to 5, 4, and 3, respectively. If $P_{UAT} = 0.3$ the probability of nonescalating decreases from 0.85 to 0.77, 0.66, and 0.51 when n decreases from 6 to 5, 4, and 3, respectively. Given the toxicity rate p , one may determine a sample size by considering the probability $P_{OT}(p)$ of overlooking this toxicity by the simple formula $P_{OT}(p) = (1 - p)^n$ (58). Given $p = 0.33$ $P_{OT}(p)$ takes the values 0.09, 0.06, 0.04, 0.03, 0.02, 0.012, 0.008 for $n = 6, 7, 8, 9, 10, 11, 12$.

1.3.4 VALIDATION AND COMPARISON OF PHASE I DESIGNS

Various simulation studies were performed to compare new designs with the standard designs TER/STER or to compare the CRM with its modifications; 23,39,40,48,49,59

see also.⁴⁶ Unfortunately, the results are based on different simulation designs and different criteria of assessment such that a comparison is rather limited.

Roughly summarizing the findings of those studies it appears that the TER is inferior to the UaD-BC and the UaD-BD³⁹ in terms of the fraction of successful trials. The UaD-BD is superior to UaD-BC at least in some situations. The BCD design from the RW class and the CRM performed very similarly.⁴⁰ By examining the distribution of the occurrence of toxic dose levels, percentage of correct estimation, or percentage treated at the MTD or one level below, the single-stage UaD designs were inferior to the CRM;²³ the UaD recommended toxic levels much more often than the CRM. STER designs were inferior to some modified CRMs with respect to the percentage of correct estimation, the percentage treated at the MTD or one level below, and the percentage of patients treated with very low levels.^{48,49} The CRM and some modified CRMs needed on the average more steps than the STER, and they showed a tendency of treating more patients at levels higher than the MTD. The STER recommended more dose levels lower than the true MTD. The STER provided no efficient estimate of the MTD and did not stop at the specified percentile of the dose–response function.⁶⁰

1.4 CONDUCT AND EVALUATION OF THE TRIAL

1.4.1 STANDARD REQUIREMENTS

Before a phase I trial is initiated, the following design characteristics should be checked and defined in the study protocol: starting dose x_1 , dose levels D , prior information on the MTD, dose–toxicity model, escalation rule, sample size per dose level, stopping rule, and a rule for completion of the sample size when stopping. It is recommended to perform a simulation study before the clinical start of the trial. Examples are found among the studies mentioned in Section 1.2.4, and software has become available recently by Simon et al.⁶¹

1.4.2 DESCRIPTIVE STATISTICAL ANALYSIS

All results obtained in a phase I trial have to be reported in a descriptive statistical analysis that accounts for the dose levels. A comprehensive and transparent report of all toxicities observed in a phase I trial is an absolute must for both the producer's (drug developer) and the consumer's (patient) risks and benefits. Absolute and relative frequencies (related to the number of patients evaluable for safety) are reported for all toxicities of the CTC list by distinguishing the grading and the assessment of the relation to treatment. A description of the individual load of toxicity of each patient has been made separately using individual descriptions eventually supported by modern graphical methods of linking scatterplots for multivariate data (see Benner et al.⁶² The evaluation of the response can usually be restricted to a case by case description of all those patients who exhibit a partial or complete response.

1.4.3 MTD ESTIMATION

The estimation of the MTD has been part of most of the search designs from Section 1.3 and an estimate MTD resulted often directly through the stopping criterion. An

estimate of a standard error of the toxicity rate at the chosen dose level is impaired by the small number of cases (≤ 6) and also by the design. A general method for analyzing dose–toxicity data is the logistic regression of the Y_j on the actually applied doses $x(j)$, $j = 1, \dots, n$, of all patients treated in the trial.⁶³ This would, however, disregard any dependency of the dose–toxicity data on the design that had created the data and may therefore be biased. If the sampling is forced to choose doses below the true MTD, $\hat{MTD}$ may be shifted toward lower values.

1.4.4 PHARMACOKINETIC PHASE I DATA ANALYSIS

An often neglected but important secondary objective of a phase I trial is the assessment of the distribution in and elimination of the drug from the body. Specific parameters that describe the pharmacokinetics of the drug are the absorption and elimination rate, the drug half-life, the peak concentration, and the area under the time–drug concentration curve (AUC). Drug concentration measurements $c_i(t_r)$ of patient i at time t_r are usually obtained from blood samples (sometimes combined with measurements from urine samples) taken regularly during medication. These data are to be analyzed using pharmacokinetic models.⁶⁴ One- and two-compartment models have been used to estimate the pharmacokinetic characteristics often in a two-step approach: first for the individual kinetic of each patient, then for the patient population using population kinetic models. In practice, statistical methodology for pharmacokinetic data analysis is based primarily on nonlinear curve fitting using least squares methods or their extensions.⁶⁵ More on statistical issues in the study of pharmacokinetics in clinical oncology is elaborated in the article of Rosner et al.⁶⁶ in this volume.

Subsequent to the criticism of traditional methods for requiring too many steps before reaching the MTD, pharmacokinetic information to reduce this number was suggested.^{25,67} Therefore, a pharmacokinetically guided dose escalation (PGDE) was proposed,^{25,68} based on the equivalence of drug blood levels in mouse and man and on the pharmacodynamic hypothesis that equal toxicity is caused by equal drug plasma levels. It postulates that the DLT is determined by plasma drug concentrations and that AUC is a measure that holds across species.⁶⁷ The AUC calculated at the MTD for humans was found to be fairly equal to the AUC for mice if calculated at the LD_{10} (in mg/m^2 equivalents, $MELD_{10}$). Therefore, $AUC(LD_{10, \text{mouse}})$ was considered as a target AUC, and a ratio

$$F = \frac{AUC(LD_{10, \text{mouse}})}{AUC(StartingDose_{\text{human}})} \quad (1.13)$$

was used to define a range of dose escalation. One tenth of $MELD_{10}$ is usually taken as the starting dose x_1 . Two variants have been proposed. In the square root method, the first step from the starting dose x_1 to the next dose x_2 is equal to the geometric mean between x_1 and the target dose $F x_1$, i.e., $x_2 = \sqrt{x_1 F x_1} = x_1 \sqrt{F}$. Subsequent dose escalation continues with the MFDE. In the extended factors of two method, the first steps are achieved by a doubling dose scheme as long as 40% of F has not been

attained. Then one continues with the MFDE. Potential problems and pitfalls were discussed in Ref. 69. Although usage and further evaluation were encouraged and guidelines for its conduct were proposed, the PGDE has not often been used. For a more recent discussion and appraisal see Newell.⁷⁰

1.4.5 DOSE TITRATION AND TOXICITY-RESPONSE APPROACH

Individual dose adjustment has been discussed repeatedly as a possible extension of the design of a phase I trial^{6,11,12,61} if a sufficient time has elapsed after the last treatment course such that any existing or lately occurring toxicity could have been observed before an elevated dose is applied. It was recommended that patients escalated to a certain level are accompanied by “fresh” patients at that level to allow the assessment of cumulative toxic effects or accumulating tolerance. When planning intraindividual dose escalation, one should weigh the advantages of dose increase and faster escalation in the patient population against the risks of cumulative toxicity in the individual patients. Further consideration should be given to the development of tolerance in some patients and the risk of treating new patients at too high levels.¹² The STER was modified as in⁶¹ by a dose titration design where two intraindividual strategies were considered: no intraindividual dose escalation and only deescalation by one level per course in case of DLT or UAT.

A design involving both dose-finding and evaluation of safety and efficacy in one early phase I/II trial was proposed by Thall and Russell⁷¹ and extended by Thall, Estey, and Sung.⁷² The goal was to define a dose satisfying both safety and efficacy requirements, to stop early when it was likely that no such dose could be found, and to continue if there was enough chance to find one. This approach combines toxicity and response into one variable, making it impossible to estimate the effect of dose upon each outcome separately. Braun⁷³ suggested a model in which for a given dose the probability of toxicity is parameterized separately from the probability of response. This approach is called bivariate CRM (bCRM). The two outcomes are then combined into a joint likelihood model that contains an additional parameter for the within-subject association of both outcomes, and the Bayesian methods of CRM can be used to estimate how dose relates to both toxicity and response.

1.4.6 REGULATIONS BY GOOD CLINICAL PRACTICE AND INTERNATIONAL CONFERENCE ON HARMONIZATION (ICH)

General guidelines for obtaining and use of dose–response information for drug registration have been given in ICH Tripartite Guideline, Dose–Response Information to Support Drug Registration Recommended for Adoption Step 4 on 10, March 1994, which covers to some extent studies with patients suffering from life-threatening diseases such as cancer. Different from standard clinical trials protocols, a phase I trial protocol may contain specifications on DLT and MTD, dose levels, and dose escalation design, and the number of patients per dose level.

According to ICH Harmonized Tripartite Guideline, General Consideration for Clinical Trials, Recommended for Adoption at Step 4 on 17 July 1997,¹⁶ the definition of pharmacokinetics and pharmacodynamics is seen as a major issue and the

study of activity or potential therapeutic benefit as a preliminary and secondary objective.

Rules for the planning and conduct of phase I trials are specifically addressed by the European Agency for the Evaluation of Medicinal Products and its Note for Guidance on *Evaluation of Anticancer Medicinal (revised version from 31 May 2001)* CPMP/EWP/205/95 (see www.tga.health.gov.au/docs/pdf/euguide/ewp/020595enrev1.pdf). Primary objectives there are laid down as determination of the MTD, characterization of frequent 'side effects' of the agent, and their dose response parameters and determination of relevant main pharmacokinetic parameters.

1.5 OPEN PROBLEMS

Different from dose-finding in bioassays, phase I trials are smaller in sample sizes, take longer for observing the endpoint, and are strictly constrained by the ethical requirement in choosing doses conservatively. Therefore, standard approaches such as unrestricted UaD and SAM designs are not applicable. Further complexity arises by population heterogeneity, subjectivity in judging toxicity, and censoring because of early drop out. Phase I trials are not completely restricted to new drugs but may be conducted also for new schedules or for new formulations of drugs in use or for drug combinations (see Ref. 74).

A comprehensive comparison of all the methods proposed is, despite a number of simulation results (mentioned in Section 3.2.5), not available at present. All previous attempts to validate a design and to compare it with others have to use simulations because a trial can be performed only once with one escalation scheme and it can not be reperformed. Therefore, all comparisons can only be as valid as the setup of the simulation study is able to cover clinical reality and complexity. There is evidence from simulations and from clinical experience that the standard designs are too conservative in terms of underdosing and needless prolongation. At the same time, the pure Bayesian and the CRM rules have deficits because their optimality is connected with treating one patient after the other. This may prolong a trial even more and risk treating too many patients at toxic levels.⁴⁸ This dilemma has motivated a large number of modifications both of the STER and the CRM (lowering the starting dose and restricting the dose escalation to one step to find a way between the conservatism and the anticonservatism). Obviously more practical experience is still needed to bring the Bayesian methods and the CRM into clinical practice.

One has to be reminded that efficacy is not in accordance with the primary aim of a phase I study, which is dose finding. The percentage of patients entered into phase I trials who benefited from phase I treatment has rarely been estimated in serious studies and may be very low; see the 3-year review⁷⁵ of 23 trials of the M.D. Anderson Cancer Center from 1991 to 1993. Therefore, it would be unrealistic to expect a therapeutic benefit even at higher dose levels for most patients in phase I studies.

Phase I trials are weak in terms of a generalization to a larger patient population because only a small number of selected patients is treated under some special circumstances by specialists. Therefore, drug development programs implement usually

more than one phase I trial for the same drug. This poses, however, ethical concerns about the premises of the phase I being the trial under which the first-time patients are treated. The repeat of phase I trials has not found much methodological consideration in the past.

Given multicentricity and enough patients per center, some restricted type of randomization may be feasible and should be considered, for example, randomly choosing the center for the first patient for the next dose level. Further improvement was put into the perspective of increasing the sample size of phase I trials and so increasing the information.¹³ The definition of the toxicity criteria and their assessment rules are mostly beyond the statistical methodology but are perhaps more crucial than any other means during the conduct of trial. Care must be taken that no changes in the assessment of toxicity occur progressively during the trial. The recent changes of the CTC criteria from version 2 to version 3 may pose some challenges on the comparability of study results.

Statistical estimation of the MTD has been restricted above to basic procedures leaving aside the question of estimation bias when neglecting the design. Babb et al.⁴³ noted that they can estimate the MTD using a different prior than used in the design and refer to work of others who suggested to use a Bayesian scheme for design and a maximum likelihood estimate of the MTD. Laboratory toxicity is, however, available on a quantitative scale, and this information could be used for the estimation of an MTD.⁷⁶ Mick and Ratain⁷⁷ proposed therefore $\ln(y) = a_0 + a_1 x + a_2 \ln(z)$ as dose-toxicity model where y was a quantitative measure of toxicity (nadir of white blood cell count) and where z was a covariate (pretreatment white blood cell count). Further research and practical application are needed to corroborate the findings. A multigrade dose escalation scheme was proposed in⁷⁸ that allows escalation as well as reduction of dosage using knowledge of the grade of toxicity. They showed that the multigrade design was superior to the standard design and that it could compete successfully with the two-stage designs. But it was not uniformly better. The authors also admitted that a multigrade design is harder to comprehend and use, and to these authors' knowledge that design has never been used in practice, although it may have influenced recent research in titration designs.

ACKNOWLEDGMENTS

This work could not have been done without the long-standing extremely fruitful cooperation with the Phase I/II Study Group of the AIO in the German Cancer Society, the work done with Wolfgang Queißer, Axel Hanauske, Heiner Fiebig, and the more recent work for the Biometric Center of the Central European Society for Anticancer Drug Research (CESAR) guided by Christian Ditttrich and Bernhard Keppler. We are indebted also for the support and statistical input from the Biometry Departments of Martin Schumacher (Freiburg) and Michael Schemper (Vienna), Harald Heinzl (Vienna), and our next-door colleague, Annette Kopp-Schneider. For technical assistance we have to thank Gudrun Friedrich for analyses and Regina Grunert and Renate Rausch for typing and for the bibliography. Finally, we would like to thank John Crowley for all the help and encouragement.

REFERENCES

1. Simon RM. A decade of progress in statistical methodology for clinical trials. *Stat. Med.* 1991;10:1789–1817.
2. World Medical Association. Declaration of Helsinki, (<http://www.wma.net/e/policy/b3.htm>).
3. Carter SK. Study design principles for the clinical evaluation of new drugs as developed by the chemotherapy programme of the National Cancer Institute. In: Staquet MJ, ed. *The Design of Clinical Trials in Cancer Therapy*. Brussels: Editions Scient. Europ., 1973:242–289.
4. Schwartzmann G, Wanders J, Koier IJ, Franklin HR, Dalesio O, Hornstra HW, et al. EORTC New Drug Development Office coordinating and monitoring programme for phase I and II trials with new anticancer agents. *Eur. J. Cancer.* 1991;27:1162–1168.
5. Spreafico F, Edelstein MB, Lelieveld P. Experimental bases for drug selection. In: Buyse ME, Staquet MJ, Sylvester RJ, eds. *Cancer Clinical Trials; Methods and Practice*. Oxford: Oxford Univ. Press, 1984:193–209.
6. Carter SK, Bakowski MT, Hellmann K. Clinical trials in cancer chemotherapy. In: Carter SK, Bakowski MT, Hellmann K, eds. *Chemotherapy of Cancer*. 3rd ed. New York: John Wiley, 1987:29–31.
7. EORTC New Drug Development Committee. EORTC guidelines for phase I trials with single agents in adults. *Eur. J. Cancer Clin. Oncol.* 1985;21:1005–1007.
8. Bodey GP, Legha SS. The phase I study: general objectives, methods, and evaluation. In: Muggia FM, Rozencweig M, eds. *Clinical Evaluation of Antitumor Therapy*. Nijhoff, The Netherlands: Dordrecht, 1987:153–174.
9. Leventhal BG, Wittes RE. Phase I trials. In: Leventhal BG, Wittes RE, eds. *Research Methods in Clinical Oncology*. New York: Raven Press, 1988:41–59.
10. Schneiderman MA. Mouse to man: statistical problems in bringing a drug to clinical trial. In: *Proc. 5th Berkeley Symp. Math. Statist. Prob.* 4th ed. Berkeley: Univ. Calif. Press, 1967:855–866.
11. Von Hoff DD, Kuhn J, Clark GM. Design and conduct of phase I trials. In: Buyse ME, Staquet MJ, Sylvester RJ, eds. *Cancer Clinical Trials; Methods and Practice*. Oxford: Oxford Univ. Press, 1984:210–220.
12. Ratain MJ, Mick R, Schilsky RL, Siegler M. Statistical and ethical issues in the design and conduct of phase I and II clinical trials of new anticancer agents. *J. Nat. Cancer Inst.* 1993;85:1637–1643.
13. Christian MC, Korn EL. The limited precision of phase I trials. *J. Nat. Cancer Inst.* 1994;86:1662–1663.
14. Kerr DJ. Phase I clinical trials: adapting methodology to face new challenges. *Ann. Oncol.* 1994;5:S67–S70.
15. Pinkel D. The use of body surface area as a criterion of drug dosage in cancer chemotherapy. *Cancer Res.* 1958;18:853–856.
16. National Cancer Institute (NCI). Cancer Therapy Evaluation Program, Common Terminology Criteria for Adverse Events. Version 3.0. DCTD, NCI, NIH, DHHS March 31, 2003 (<http://www.ctep.cancer.gov>) Publication date June 10, 2003.
17. World Health Organization (WHO). *WHO Handbook for Reporting Results of Cancer Treatment*. WHO Offset Publication No.48. Genf: WHO, 1979.
18. Kreuser ED, Fiebig HH, Scheulen ME, Hanauske AR, Keppler BK, Mross KB, et al. Standard operating procedures and organization of German phase I, II, and III study groups, New Drug Development Group (AWO), and Study Group of Pharmacology

- in Oncology and Hematology (APOH) of the Association for Medical Oncology (AIO) of the German Cancer Society. *Onkol.* 1998;21 (Suppl.3).
19. Mick R, Lane N, Daugherty C, Ratain MJ. Physician-determined patient risk of toxic effects: impact on enrollment and decision making in phase I trials. *J. Nat. Cancer Inst.* 1994;86:1685–1693.
 20. Franklin HR, Simonetti GPC, Dubbelman AC, Ten Bokkel Huinink WW, Taal BG, Wigbout G, et al. Toxicity grading systems: a comparison between the WHO scoring system and the Common Toxicity Criteria when used for nausea and vomiting. *Ann. Oncol.* 1994;5:113–117.
 21. Brundage MD, Pater JL, Zee B. Assessing the reliability of two toxicity scales: implications for interpreting toxicity data. *J. Nat. Cancer Inst.* 1993;85:1138–1148.
 22. Morgan BJT. *Analysis of Quantal Response Data*. London: Chapman & Hall, 1992.
 23. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometr.* 1990;46:33–48.
 24. Edler L. Statistical requirements of phase I studies. *Onkologie.* 1990;13:90–95.
 25. Collins JM, Zaharko DS, Dedrick RL, Chabner BA. Potential roles for preclinical pharmacology in phase I clinical trials. *Cancer Treat. Rep.* 1986;70:73–80.
 26. Goldsmith MA, Slavik M, Carter SK. Quantitative prediction of drug toxicity in humans from toxicology in small and large animals. *Cancer Res.* 1975;35:1354–1364.
 27. Hansen HH. Clinical experience with 1-(2-chloroethyl)-3-cyclohexyl-1-nitrosourea (CCNU, NSC-79037). *Proc. Am. Assoc. Cancer Res.* 1970;11:87.
 28. Muggia FM. Phase I study of 4'demethyl-epipodophyllotoxin- D-thenylidene glycoside (PTG, NSC-122819). *Proc. Am. Assoc. Cancer Res.* 1970;11:58.
 29. Carter SK. Clinical trials in cancer chemotherapy. *Cancer* 1977;40:544–557.
 30. Hansen HH, Selawry OS, Muggia FM, Walker MD. Clinical studies with 1-(2-chloroethyl)-3-cyclohexyl-1-nitrosourea (NSC79037). *Cancer Res.* 1971;31:223–227.
 31. Bellman RE. Topics in pharmacokinetics, III: repeated dosage and impulse control. *Math. Biosci.* 1971;12:1–5.
 32. Bellman RE. *Dynamic Programming*. Princeton: Princeton Univ. Press, 1962.
 33. DeVita VT, Carbone PP, Owens AH, Gold GL, Krant MJ, Emonson J. Clinical trials with 1,3-Bis(2-chlorethyl)-1-nitrosourea, NSC 409962. *Cancer Treat. Rep.* 1965; 25:1876–1881.
 34. Wijnand HP. Updates of bioequivalence programs (including statistical power approximated by student's t). *Comp. Meth. Progr. Biomed.* 1994; 42:275–281.
 35. Penta JS, Rozenzweig M, Guarino AM, Muggia FM. Mouse and large-animal toxicology studies of twelve antitumor agents: relevance to starting dose for phase I clinical trials. *Cancer Chemother. Pharmacol.* 1979;3:97–101.
 36. Rozenzweig M, Von Hoff DD, Staquet MJ, Schein PS, Penta JS, Goldin A, et al. Animal toxicology for early clinical trials with anticancer agents. *Cancer Clin. Trials* 1981;4:21–28.
 37. Willson JKV, Fisher PH, Tutsch K, Alberti D, Simon K, Hamilton RD, Bruggink J, Koeller JM, Tormey DC, Earhardt RH, Ranhosky A, Trump DL. Phase I clinical trial of a combination of dipyrindamole and acivicin based upon inhibition of nucleoside salvage. *Cancer Res.* 1988;48:5585–5590.
 38. Storer BE. Choosing a phase I design. In: Crowley J, ed. *Handbook of Statistics in Clinical Oncology*. New York, Marcel Dekker, 2005: This volume.
 39. Storer BE. Design and analysis of phase I clinical trials. *Biometr.* 1989;45:925–937.
 40. Durham SD, Flournoy N, Rosenberger WF. A random walk rule for phase I clinical trials. *Biometr.* 1997;53:745–760.

41. Dixon WJ, Mood AM. A method for obtaining and analyzing sensitivity data. *J. Am. Statist. Assoc.* 1948;43:109–126.
42. Robbins H, Monro S. A stochastic approximation method. *Ann. Math. Stat.* 1951; 29:400–407.
43. Babb J, Rogatko A, Zacks S. Cancer phase I clinical trials: efficient dose escalation with overdose control. *Stat. Med.* 1998;17:1103–1120.
44. Whitehead J, Brunier H. Bayesian decision procedures for dose determining experiments. *Stat. Med.* 1995;14:885–893.
45. Gatsonis C, Greenhouse JB. Bayesian methods for phase I clinical trials. *Stat. Med.* 1992;11:1377–1389.
46. O'Quigley J. Phase I and Phase I/II dose finding algorithms using continual reassessment method. In: Crowley J, ed. *Handbook of Statistics in Clinical Oncology*. New York, Marcel Dekker, 2005: This volume.
47. Shen LZ, O'Quigley J. Consistency of continual reassessment method under model misspecification. *Biometrika.* 1996;83:395–405.
48. Korn EL, Midthune D, Chen TT, Rubinstein LV, Christian MC, Simon RM. A comparison of two phase I trial designs. *Stat. Med.* 1994;13:1799–1806.
49. Faries D. Practical modifications of the continual reassessment method for phase I clinical trials. *J. Biopharm. Stat.* 1994;4:147–164.
50. Moller S. An extension of the continual reassessment methods using a preliminary up-and-down design in a dose finding study in cancer patients, in order to investigate a greater range of doses. *Stat. Med.* 1995;14:911–922.
51. Goodman SN, Zahurak ML, Piantadosi S. Some practical improvements in the continual reassessment method for phase I studies. *Stat. Med.* 1995;14:1149–1161.
52. Ahn C. An evaluation of phase I cancer clinical trial designs. *Stat. Med.* 1998; 17:1537–1549.
53. Hanauske AR, Edler L. New clinical trial designs for phase I studies in hematology and oncology: principles and practice of the continual reassessment model. *Onkologie* 1996;19:404–409.
54. O'Quigley J, Reiner E. A stopping rule for the continual reassessment method. *Biometrika* 1998;85:741–748.
55. Heyd JM, Carlin BP. Adaptive design improvements in the continual reassessment method for phase I studies. *Stat. Med.* 1999;18:1307–1321.
56. O'Quigley J. Continual reassessment designs with early termination. *Biostat.* 2002; 3:87–99.
57. Rademaker AW. Sample sizes in phase I toxicity studies. *ASA Proc. Biopharm. Sect.* 1989;137–141.
58. Geller NL. Design of phase I and II clinical trials in cancer: a statistician's view. *Cancer Invest.* 1984;2:483–491.
59. 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.
60. Storer BE. Small-sample confidence sets for the MTD in a phase I clinical trial. *Biometr.* 1993;49:1117–1125.
61. Simon RM, Freidlin B, Rubinstein LV, Arbuck SG, Collins J, Christian MC. Accelerated titration designs for phase I clinical trials in oncology. *J. Nat. Cancer Inst.* 1997;89:1138–1147.
62. Benner A, Edler L, Hartung G. SPLUS support for analysis and design of phase I trials in clinical oncology. In: Millard S, Krause A, eds. *SPLUS in Pharmaceutical Industry*. New York: Springer, 2000.
63. Finney DJ. *Statistical Method in Biological Assay*. London: C. Griffin, 1978.

64. Gibaldi M, Perrier D. *Pharmacokinetics*. New York: M. Dekker, 1982.
65. Edler L. Computational statistics for pharmacokinetic data analysis. In: Payne R, Green P, eds. *COMPSTAT. Proceedings in Computational Statistics*. Heidelberg: Physica, 1998:281–286.
66. Rosner GL, Mueller P, Lunagomez S, Thomson PA. Pharmacokinetics in clinical oncology: Statistical Issues. In: Crowley J, ed. *Handbook of Statistics in Clinical Oncology*. New York, Marcel Dekker, 2005: This volume.
67. Collins JM. Pharmacology and drug development. *J. Nat. Cancer Inst.* 1988; 80: 790–792.
68. Collins JM, Grieshaber CK, Chabner BA. Pharmacologically guided phase I clinical trials based upon preclinical drug development. *J. Nat. Cancer Inst.* 1990; 82:1321–1326.
69. EORTC Pharmacokinetics and Metabolism Group. Pharmacokinetically guided dose escalation in phase I clinical trials: commentary and proposed guidelines. *Eur. J. Cancer Clin. Oncol.* 1987;23:1083–1087.
70. Newell DR. Pharmacologically based phase I trials in cancer chemotherapy. *New Drug Ther.* 1994;8:257–275.
71. Thall PF, Russell KE. A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometr.* 1998;54:251–264.
72. Thall PF, Estey EH, Sung HG. A new statistical method for dose-finding based on efficacy and toxicity in early phase clinical trials. *Investig. New Drugs* 1999; 17:251–264.
73. Braun TM. The bivariate continual reassessment method: extending the CRM to phase I trials of two competing outcomes. *Controlled Clinical Trials* 2002;23:240–256.
74. Korn EL, Simon RM. Using the tolerable-dose diagram in the design of phase I combination chemotherapy trials. *J. Clin. Oncol.* 1993;11:794–801.
75. Smith TL, Lee JJ, Kantarjian HM, Legha SS, Raber MN. Design and results of phase I cancer clinical trials: three-year experience at M.D. Anderson Cancer Center. *J. Clin. Oncol.* 1996;14:287–295.
76. Egorin MJ. Phase I trials: a strategy of ongoing refinement. *J. Nat. Cancer Inst.* 1990; 82:446–447.
77. Mick R, Ratain MJ. Model-guided determination of maximum tolerated dose in phase I clinical trials: evidence for increased precision. *J. Nat. Cancer Inst.* 1993; 85:217–223.
78. Gordon NH, Willson JKV. Using toxicity grades in the design and analysis of cancer phase I clinical trials. *Stat. Med.* 1992;11:2063–2075.

2 Phase I and Phase I/II Dose Finding Algorithms Using Continual Reassessment Method

John O'Quigley

CONTENTS

Summary	32
2.1 Continual Reassessment Method	32
2.1.1 Motivation	32
2.1.2 Operating Characteristics	35
2.2 Technical Aspects.....	36
2.2.1 Model Requirements	37
2.3 Implementation	38
2.3.1 Maximum Likelihood Implementation	39
2.3.2 Two-Stage Designs.....	40
2.3.3 Grouped designs.....	42
2.3.4 Illustration	42
2.4 Statistical Properties.....	43
2.4.1 Convergence.....	43
2.4.2 Efficiency	44
2.4.3 Safety	46
2.5 More complex CRM designs	47
2.5.1 Inclusion of a Stopping Rule	47
2.5.2 Patient Heterogeneity	48
2.5.3 Pharmacokinetic Studies	50
2.5.4 Graded Toxicities	51
2.6 Combined Toxicity-Efficacy Studies	53
2.7 Bayesian Approaches	55
References	57

SUMMARY

We review the continual reassessment method (O'Quigley, Pepe and, Fisher 1990) as it is used in phase I and phase I/II clinical dose finding studies in oncology. We describe the main ideas behind the method and how to implement it in practice. Although a Bayesian framework is sometimes employed, the method itself is not essentially Bayesian (O'Quigley and Shen 1996; Storer 1998). Our emphasis here is mostly on maximum likelihood estimation. For general use we recommend a class of two-stage designs in which the investigator has considerable flexibility in the first stage, followed by a second stage leaning on a CRM model and maximum likelihood implementation. For phase I/II designs, in which information on efficacy can be obtained within a comparable time frame to that on toxicity, we can use the CRM structure together with the sequential probability ratio test to construct very effective designs to locate the dose producing the greatest rate of success. The statistical efficiency of any design is discussed, and we consider an efficiency bound that is theoretical in a way analogous to the Cramer–Rao bound for an estimator and that enables us to gauge how fruitful the search for newer designs potentially could be.

Key words: clinical trial; continual reassessment method; dose escalation; dose finding studies; efficacy studies; efficiency; maximum likelihood; Phase I trial; phase I/II trial; toxicity.

2.1 CONTINUAL REASSESSMENT METHOD

This review describes the basic ideas behind the continual reassessment method (CRM) as it is used in phase I and phase I/II dose finding. We recall some important technical considerations and some key properties of the method as well as the possibility for substantial generalization; specifically, the use of graded information on toxicities, the incorporation of a stopping rule leading to further reductions in sample size, the incorporation of information on patient heterogeneity, the incorporation of pharmacokinetics, and the possibility of modeling within patient dose escalation. In its most classical setting the CRM is used to identify the maximum tolerated dose (MTD) where only information on toxicities is used. More recently (O'Quigley, Hughes, and Fenton 2001; Ivanova 2003), for studies in which we observe information on both toxicity and efficacy, the method can be used to identify a level having the greatest rate of overall success.

2.1.1 MOTIVATION

Storer (1989) made explicit the goal of a phase I dose finding study in chronic illness such as cancer as being the identification of some dose corresponding to an acceptable rate of undesirable side effects, usually called toxicities. This would be the context for cytotoxic drugs in which we might view toxicity as a surrogate for longer term efficacy. For phase I/II studies we observe in a similar time frame both toxicity and some measure of effect. For these studies the goal is usually to maximize the overall

success rate. O'Quigley, Pepe, and Fisher (1990) argued that for a phase I study, in addition to the goal of targeting some percentile, an acceptable design should aim to incorporate the following restrictions:

1. We should minimize the number of undertreated patients, i.e., patients treated at unacceptably low dose levels.
2. We should minimize the number of patients treated at unacceptably high dose levels.
3. We should minimize the number of patients needed to complete the study (efficiency).
4. The method should respond quickly to inevitable errors in initial guesses, rapidly escalating in the absence of indication of drug activity (toxicity) and rapidly deescalating in the presence of unacceptably high levels of observed toxicity.

Before describing just how the CRM meets the above requirements, let us first look at the requirements themselves in the context of phase I cancer dose finding studies.

Most phase I cancer clinical trials are carried out on patients for whom all currently available therapies have failed. There will always be hope in the therapeutic potential of the new experimental treatment, but such hope often will be tempered by the almost inevitable life-threatening toxicity accompanying the treatment. Given that candidates for these trials have no other options concerning treatment, their inclusion appears contingent upon maintaining some acceptable degree of control over the toxic side effects as well as trying to maximize treatment efficacy (which in the absence of information on efficacy itself translates as dose). Too high a dose, while offering in general better hope for treatment effect, will be accompanied by too high a probability of encountering unacceptable toxicity. Too low a dose, while avoiding this risk, may offer too little chance of seeing any benefit at all.

Given this context, the first two of the listed requirements appear immediate. The third requirement, a concern for all types of clinical studies, becomes of paramount importance here where very small sample sizes are inevitable. This is because of the understandable desire to proceed quickly with a potentially promising treatment to the phase II stage. At the phase II stage the probability of observing treatment efficacy is almost certainly higher than that for the phase I population of patients. We have to do the very best we can with the relatively few patients available, and the statistician involved in such studies also should provide some idea as to the error of our estimates, translating the uncertainty of our final recommendations based on such small samples. The fourth requirement is not an independent requirement and can be viewed as a partial reexpression of requirements 1 and 2.

Taken together, the requirements point toward a method where we converge quickly to the correct level, the correct level being defined as the one having a probability of toxicity as close as possible to some value θ . The value θ is chosen by the investigator such that he or she considers probabilities of toxicity higher than θ to be unacceptably high, while those lower than θ are unacceptably low in that they indirectly, indicate, the likelihood of too weak an antitumor effect.

Figure 2.1 illustrates the comparative behavior of CRM with a fixed sample up-and-down design in which level 7 is the correct level. How does CRM work? The essential idea is similar to that of stochastic approximation, the main differences being the use of a nonlinear underparameterized model belonging to a particular class of models and a small number of discrete dose levels rather than a continuum. Patients enter sequentially. The working dose-toxicity curve, taken from the CRM class (described below), is refitted after each inclusion. The curve is then inverted to identify which of the available levels has an associated estimated probability as close as we can get to the targeted acceptable toxicity level. The next patient is then treated at this level. The cycle is continued until a fixed number of subjects have been

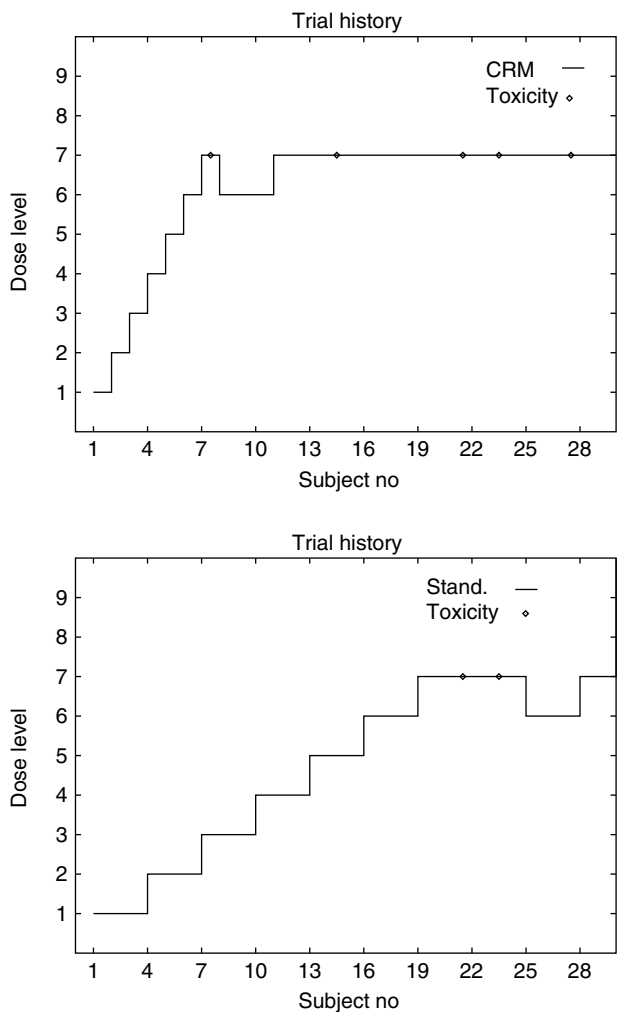


FIGURE 2.1 Typical trial histories: CRM (above) and standard design (below).

treated or until we apply some stopping rule (see Section 2.5). Typical behavior is that shown in [Figure 2.1](#).

2.1.2 OPERATING CHARACTERISTICS

A large sample study (Shen and O'Quigley 1996) showed that under some broad conditions the level to which a CRM design converges will indeed be the closest to the target. As pointed out by Storer (1998), large sample properties themselves will not be wholly convincing because in practice, we are inevitably faced with small to moderate sample sizes. Nonetheless, if any scheme fails to meet such basic statistical criteria as large sample convergence, we need investigate with great care its finite sample properties. The tool to use here is mostly that of simulation, although for the standard up- and -down schemes, the theory of Markov chains enables us to carry out exact probabilistic calculations (Storer 1993; Reiner, Paoletti, and O'Quigley 1998).

Whether Bayesian or likelihood based, once the scheme is under way (i.e., the likelihood is nonmonotone), then it is readily shown that a nontoxicity always points in the direction of higher levels and a toxicity in the direction of lower levels, the absolute value of the change diminishing with the number of included patients. For nonmonotone likelihood it is impossible to be at some level, observe a toxicity, and then for the model to recommend a higher level as claimed by some authors (see Section 2.6.2) unless pushed in such a direction by a strong prior. Furthermore, when targeting lower percentiles such as 0.2, it can be calculated, and follows our intuition, that a toxicity occurring with a frequency factor of 4 less than that for the nontoxicities will have a much greater impact on the likelihood or posterior density. This translates directly into an operating characteristic whereby model based escalation is relatively cautious and deescalation more rapid, particularly early on where little information is available. In the model and examples of O'Quigley, Pepe, and Fisher (1990), dose levels could never be skipped when escalating. However, if the first patient, treated at level 3, suffered a toxic side effect, the method skipped when deescalating, recommending level 1 for the subsequent two entered patients before, assuming no further toxicities were seen, escalating to level 2.

Simulations in O'Quigley, Pepe, and Fisher (1990), O'Quigley and Chevret (1991), Goodman et al. (1995), Korn et al. (1994), and O'Quigley (1999) show the operating characteristics of CRM to be good in terms of accuracy of final recommendation while simultaneously minimizing the numbers of overtreated and undertreated patients. However, violation of the model requirements and allocation principle of CRM described in the following section can have a negative, possibly disastrous, effect on operating characteristics (Chevret 1993) resulting in certain cases in never recommending the correct level, a performance worse than we achieve by random guessing. Korn et al. (1994) made this mistake, and their results, already favorable to CRM, would have been yet more favorable had a model not violating the basic requirements been used (O'Quigley 1999). Both Faries (1994) and Moller (1995) assigned to levels other than those indicated by the model, leading to large skips in the dose allocation, in one case skipping 9 levels after the inclusion of a single patient. We return to this in the discussion.

2.2 TECHNICAL ASPECTS

The aim of CRM is to locate the most appropriate dose, the so-called target dose, the precise definition of which is provided below. This dose is taken from some given range of available doses. The problem of dose spacing for single drug combinations, often addressed via a modified Fibonacci design, is beyond the scope of CRM. The need to add doses may arise in practice when the toxicity frequency is deemed too low at one level but too toxic at the next higher level. CRM can help with this affirmation, but as far as extrapolation or interpolation of dose is concerned, the relevant insights will come from pharmacokinetics. For our purposes we assume that we have available k fixed doses; $d_1, \dots, d_k$. These doses are not necessarily ordered in terms of the d_i themselves, in particular since each d_i may be a vector, being a combination of different treatments, but rather in terms of the probability $R(d_i)$ of encountering toxicity at each dose d_i . It is important to be careful at this point because confusion can arise over the notation. The d_i , often multidimensional, describe the actual doses or combinations of doses being used. We assume monotonicity, and we take monotonicity to mean that the dose levels, equally well identified by their integer subscripts i ($i = 1, \dots, k$), are ordered whereby the probability of toxicity at level i is greater than that at level i' whenever $i > i'$. The monotonicity requirement or the assumption that we can so order our available dose levels is thus important. Currently all the dose information required to run a CRM trial is contained in the dose levels. Without wishing to preclude the possibility of exploiting information contained in the doses d_i and not in the dose levels i , at present we lose no information when we replace d_i by i .

The actual amount of drug, therefore (so many mg/m², say), is typically not used. For a single agent trial (see Moller 1995), it is in principle possible to work with the actual dose. We do not advise this because it removes, without operational advantages, some of our modeling flexibility. For multidrug or treatment combination studies, there is no obvious univariate measure. We work instead with some conceptual dose, increasing when one of the constituent ingredients increases and, under our monotonicity assumption, translating itself as an increase in the probability of a toxic reaction. Choosing the dose levels amounts to selecting levels (treatment combinations) such that the lowest level will have an associated toxic probability less than the target and the highest level possibly will be close to or higher than the target.

The most appropriate dose, the target dose, is that dose having an associated probability of toxicity as close as we can get to the target "acceptable" toxicity θ . Values for the target toxicity level θ might typically be 0.2, 0.25, 0.3, 0.35, although there are studies in which this can be as high as 0.4 (Moller 1995). The value depends on the context and the nature of the toxic side effects.

The dose for the j th entered patient, X_j can be viewed as random taking values x_j , most often discrete, in which case $x_j \in \{d_1, \dots, d_k\}$, but possibly continuous, where $X_j = x$; $x \in \mathbb{R}^+$. In light of the previous two paragraphs, we can if desired entirely suppress the notion of dose and retain only information pertaining to dose level. This is all we need, and we may prefer to write $x_j \in \{1, \dots, k\}$. Let Y_j be a binary random variable (0, 1) where 1 denotes severe toxic response for the j th

entered patient ($j = 1, \dots, n$). We model $R(x_j)$, the true probability of toxic response at $X_j = x_j$; $x_j \in \{d_1, \dots, d_k\}$; or $x_j \in \{1, \dots, k\}$ via

$$R(x_j) = \Pr(Y_j = 1 | X_j = x_j) = E(Y_j | x_j) = \psi(x_j, a)$$

for some one parameter model $\psi(x_j, a)$.

For the most common case of a single homogeneous group of patients, we are obliged to work with an underparameterized model, notably a one parameter model. Although a two parameter model may appear more flexible, the sequential nature of CRM, together with its aim to put the included patients at a single correct level, means that we will not obtain information needed to fit two parameters. We are close to something like nonidentifiability. A likelihood procedure will be unstable and may even break down, whereas a two parameter fully Bayesian approach (O'Quigley, Pepe, and Fisher 1990; Gatsonis and Greenhouse 1992; Whitehead and Williamson 1998) may work initially, although somewhat artificially, but behave erratically as sample size increases (see also Section 2.6). This is true even when starting out at a low or the lowest level, initially working with an up- and -down design for early escalation, before a CRM model is applied. Indeed, any design that ultimately concentrates all patients from a single group on some given level can fit no more than a single parameter without running into problems of consistency.

2.2.1 MODEL REQUIREMENTS

The restrictions on $\psi(x, a)$ were described by O'Quigley, Pepe, and Fisher (1990). For given fixed x , we require that $\psi(x, a)$ be strictly monotonic in a . For fixed a we require that $\psi(x, a)$ be monotonic increasing in x or, in the usual case of discrete dose levels d_i $i = 1, \dots, k$, that $\psi(d_i, a) > \psi(d_m, a)$ whenever $i > m$. The true probability of toxicity at x (i.e., whatever treatment combination has been coded by x) is given by $R(x)$, and we require that for the specific doses under study ($d_1, \dots, d_k$) there exists values of a , say $a_1, \dots, a_k$, such that $\psi(d_i, a_i) = R(d_i)$, ($i = 1, \dots, k$). In other words, our one parameter model has to be rich enough to model the true probability of toxicity at any given level. We call this a working model because we do not anticipate a single value of a to work precisely at every level, i.e., we do not anticipate $a_1 = a_2 = \dots = a_k = a$. Many choices are possible. We have obtained excellent results with the simple choice:

$$\psi(d_i, a) = \alpha_i^a, \quad (i = 1, \dots, k) \quad (2.1)$$

where $0 < \alpha_1 < \dots < \alpha_k < 1$ and $0 < \alpha < \infty$. For the six levels studied in the simulations by O'Quigley, Pepe, and Fisher (1990), the working model had $\alpha_1 = 0.05$, $\alpha_2 = 0.10$, $\alpha_3 = 0.20$, $\alpha_4 = 0.30$, $\alpha_5 = 0.50$, and $\alpha_6 = 0.70$. In that paper, this was expressed a little differently in terms of conceptual dose d_i where $d_1 = -1.47$, $d_2 = -1.10$, $d_3 = -0.69$, $d_4 = -0.42$, $d_5 = -0.0$, and $d_6 = -0.42$ obtained from a model in which

$$\alpha_i = (\tanh d_i + 1)/2 \quad (i = 1, \dots, k) \quad (2.2)$$

The above “tanh” model was first introduced in this context by O’Quigley, Fisher, and Pepe (1990), the idea being that $\tanh(x)$ increases monotonically from 0 to 1 as x increases from $-\infty$ to ∞ . This extra generality is not usually needed because attention is focused on the few fixed d_i . Note that, at least as far as maximum likelihood estimation is concerned (see Section 2.3.1), working with model (2.1) is equivalent to working with a model in which $\alpha_i (i = 1, \dots, k)$ is replaced by $\alpha_i^* (i = 1, \dots, k)$ where $\alpha_i^* = \alpha_i^m$ for any real $m > 0$. Thus we cannot really attach any concrete meaning to the α_i . However, the spacing between adjacent α_i will impact operating characteristics. Working with real doses corresponds to using some fixed dose spacing, although not necessarily one with nice properties. The spacings chosen here have proved satisfactory in terms of performance across a broad range of situations. An investigation into how to choose the α_i with the specific aim of improving certain aspects of performance has yet to be carried out.

Some obvious choices for a model can fail the above conditions leading to potentially poor operating characteristics. The one parameter logistic model, $\psi(x, a) = w/(1 + w)$, in which b is fixed and where $w = \exp(b + ax)$ can be seen to fail the above requirements (Shen and O’Quigley 1996). In contrast, the less intuitive model obtained by redefining w so that $w = \exp(a + bx)$, $b \neq 0$, belongs to the CRM class.

2.3 IMPLEMENTATION

Once a model has been chosen and we have data in the form of the set $\Omega_j = \{y_1, x_1, \dots, y_j, x_j\}$, from the outcomes of the first j experiments we obtain estimates $\hat{R}(d_i)$, ($i = 1, \dots, k$) of the true unknown probabilities $R(d_i)$, ($i = 1, \dots, k$) at the k dose levels (see below). The target dose level is that level having associated with it a probability of toxicity as close as we can get to θ . The dose or dose level x_j assigned to the j th included patient is such that

$$|\hat{R}(x_j) - \theta| < |\hat{R}(d_i) - \theta|, \quad (i = 1, \dots, k; x_j \neq d_i)$$

Thus x_j is the closest level to the target level in the above precise sense. Other choices of closeness could be made incorporating cost or other considerations. We could also weight the distance, for example multiply $|\hat{R}(x_j) - \theta|$ by some constant greater than 1 when $\hat{R}(x_j) > \theta$. This would favor conservatism, such a design tending to experiment more often below the target than a design without weights. Similar ideas have been pursued by Babb et al. (1998).

The estimates $\hat{R}(x_j)$ obtain from the one parameter working model. Two questions dealt with in this section arise: (1) how do we estimate $R(x_j)$ on the basis of Ω_{j-1} , (2) how do we obtain the initial data, in particular since the first entered patient or group of patients must be treated in the absence of any data based estimates of $R(x_1)$.

Even though our model is underparameterized leading us into the area of misspecified models, it turns out that standard procedures of estimation work. Some care is needed to show this and we look at this in Section 2.4. The procedures themselves are described just below. Obtaining the initial data is partially described in these same sections as well as being the subject of its own subsection, *Two-Stage Designs*.

In order to decide on the basis of available information and previous observations the appropriate level at which to treat a patient, we need some estimate of the probability of toxic response at dose level d_i , ($i = 1, \dots, k$). We would currently recommend use of the maximum likelihood estimator (O'Quigley and Shen 1996) described in 2.3.1. The Bayesian estimator developed in the original paper by O'Quigley, Pepe, and Fisher (1990) will perform very similarly unless priors are strong. The use of strong priors in the context of an underparametrized and misspecified model may require deeper study. Bayesian ideas can nonetheless be very useful in addressing more complex questions such as patient heterogeneity and intra-patient escalation. We return to this in Section 2.7.

2.3.1 MAXIMUM LIKELIHOOD IMPLEMENTATION

After the inclusion of the first j patients, the log-likelihood can be written as:

$$\mathcal{L}_j(a) = \sum_{\ell=1}^j y_{\ell} \log \psi(x_{\ell}, a) + \sum_{\ell=1}^j (1 - y_{\ell}) \log(1 - \psi(x_{\ell}, a)) \quad (2.3)$$

and is maximized at $a = \hat{a}_j$. Maximization of $\mathcal{L}_j(a)$ can easily be achieved with a Newton–Raphson algorithm or by visual inspection using some software package such as Excel. Once we have calculated $\hat{a}_j$, we can next obtain an estimate of the probability of toxicity at each dose level d_i via:

$$\hat{R}(d_i) = \psi(d_i, \hat{a}_j), \quad (i = 1, \dots, k)$$

On the basis of this formula, the dose to be given to the $(j + 1)$ th patient, x_{j+1} is determined. We also can calculate an approximate 100 $(1 - \alpha)\%$ confidence interval for $\psi(x_{j+1}, \hat{a}_j)$ as (ψ_j^-, ψ_j^+) where

$$\psi_j^- = \psi\{x_{j+1}, (\hat{a}_j + z_{1-\alpha/2}v(\hat{a}_j)^{1/2})\}, \quad \psi_j^+ = \psi\{x_{j+1}, (\hat{a}_j - z_{1-\alpha/2}v(\hat{a}_j)^{1/2})\}$$

where z_{α} is the α th percentile of a standard normal distribution and $v(\hat{a}_j)$ is an estimate of the variance of $\hat{a}_j$. For the model of equation (2.1) this turns out to be particularly simple, and we can write

$$v^{-1}(\hat{a}_j) = \sum_{\ell \leq j, y_{\ell}=0} \psi(x_{\ell}, \hat{a}_j)(\log \alpha)^2 / (1 - \psi(x_{\ell}, \hat{a}_j))^2$$

Although based on a misspecified model, these intervals turn out to be quite accurate even for sample sizes as small as 12 and thus helpful in practice (Natarajan and O'Quigley 2003).

A requirement to be able to maximize the log-likelihood on the interior of the parameter space is that we have heterogeneity among the responses, i.e., at least one toxic and one nontoxic response (Silvapulle 1981). Otherwise the likelihood is maximized on the boundary of the parameter space and our estimates of $R(d_i)$, ($i = 1, \dots, k$) are trivially either zero, one, or depending on the model we are working with, may not even be defined.

Thus the experiment is considered as not being fully underway until we have some heterogeneity in the responses. These could arise in a variety of different ways: use of the standard up- and -down approach, use of an initial Bayesian CRM as outlined below, or use of a design believed to be more appropriate by the investigator. Once we have achieved heterogeneity, the model kicks in and we continue as prescribed above (estimation–allocation). Getting the trial underway, i.e., achieving the necessary heterogeneity to carry out the above prescription, is largely arbitrary. This feature is specific to the maximum likelihood implementation and such that it may well be treated separately.

2.3.2 TWO-STAGE DESIGNS

It may be felt that we know so little before undertaking a given study that it is worthwhile splitting the design into 2 stages, an initial exploratory escalation followed by a more refined homing in on the target. Such an idea was first proposed by Storer (1989) in the context of the more classical up- and -down schemes. His idea was to enable more rapid escalation in the early part of the trial where we may be quite far from a level at which treatment activity could be anticipated. Moller (1995) was the first to use this idea in the context of CRM designs. Her idea was to allow the first stage to be based on some variant of the usual up- and -down procedures.

In the context of sequential likelihood estimation, the necessity of an initial stage was pointed out by O’Quigley and Shen (1996) because the likelihood equation fails to have a solution on the interior of the parameter space unless some heterogeneity in the responses has been observed. Their suggestion was to work with any initial scheme Bayesian CRM, or up- and -down, and for any reasonable scheme the operating characteristics appear relatively insensitive to this choice.

However, we believe there is something very natural and desirable in two-stage designs and that currently they could be taken as the designs of choice. The reason is the following; early behavior of the method in the absence of heterogeneity, i.e., lack of toxic response, appears to be rather arbitrary. A decision to escalate after inclusion of 3 patients tolerating some level or after a single patient tolerating a level or according to some Bayesian prior, however constructed, is translating directly (less directly for the Bayesian prescription) the simple desire to try a higher dose because, thus far, we have encountered no toxicity.

The use of a working model at this point, as occurs for Bayesian estimation, may be somewhat artificial, and the rate of escalation can be modified at will, albeit somewhat indirectly, by modifying our model parameters or our prior. Rather than lead the clinician into thinking that something subtle and carefully analytic is taking place, our belief is that it is preferable that he or she be involved in the design of the initial phase. Operating characteristics that do not depend on data ought to be driven by clinical rather than statistical concerns. More importantly, the initial phase of the design, in which no toxicity has yet been observed, can be made much more efficient from both the statistical and ethical angles by allowing information on toxicity grade to determine the rapidity of escalation.

The simplest example of a two-stage design would be to include an initial escalation stage that exactly replicates the old standard design, i.e., starting at the lowest

level, 3 patients are treated, and only if all 3 tolerate the dose do we escalate to a higher level. As soon as the first dose limiting toxicity (DLT) is encountered, we close the first stage and open the second stage based on CRM modeling and using all the available data. Such a scheme could be varied in many ways, for example including only a single patient at the lowest level, then two patients at the second lowest, and then as before. The following design, which has been used in practice, includes in the first stage information on the toxicity grade. There were many dose levels, and the first included patient was treated at a low level. As long as we observe very low grade toxicities, we escalate quickly, including only a single patient at each level. As soon as we encounter more serious toxicities, escalation is slowed down. Ultimately we encounter dose limiting toxicities, at which time the second stage, based on fitting a CRM model, comes fully into play. This is done by integrating this information and that obtained on all the earlier non-dose-limiting toxicities to estimate the most appropriate dose level.

It was decided to use information on low grade toxicities in the first stage of a two-stage design in order to allow rapid initial escalation because it is possible that we could be far below the target level. Specifically we define a grade severity variable $S(i)$ to be the average toxicity severity observed at dose level i , i.e., the sum of the severities at that level divided by the number of patients treated at that level (Table 2.1). The rule is to escalate providing $S(i)$ is less than 2. Furthermore, once we have included 3 patients at some level, then escalation to higher levels only occurs if each cohort of 3 patients does not experience dose limiting toxicity. This scheme means that, in practice, as long as we see only toxicities of severities coded 0 or 1 then we escalate. The first severity coded 2 necessitates a further inclusion at this same level, and anything other than a 0 severity for this inclusion would require yet a further inclusion and a non-dose-limiting toxicity before being able to escalate. This design also has the advantage that, should we be slowed down by a severe (severity 3), albeit non-dose-limiting toxicity, we retain the capability of picking up speed (in escalation) should subsequent toxicities be of low degree (0 or 1). This can be helpful in avoiding being handicapped by an outlier or an unanticipated and possibly not drug related toxicity.

Once dose limiting toxicity is encountered, this phase of the study (the initial escalation scheme) comes to a close and we proceed on the basis of CRM recommendation. Although the initial phase is closed, the information on both dose limiting and non-dose-limiting toxicities thereby obtained is used in the second stage.

TABLE 2.1
Toxicity “Grades” (Severities) for Trial

Severity	Degree of Toxicity
0	No toxicity
1	Mild toxicity (non-dose-limiting)
2	Non-mild toxicity (non-dose-limiting)
3	Severe toxicity (non-dose-limiting)
4	Dose limiting toxicity

2.3.3 GROUPED DESIGNS

O'Quigley, Pepe, and Fisher (1990) describe the situation of delayed response in which new patients become available to be included in the study while we are still awaiting the toxicity results on already entered patients. The suggestion was, in the absence of information on such recently included patients, that the logical course to take was to treat at the last recommended level. This is the level indicated by all the currently available information.

The likelihood for this situation was written down by O'Quigley, Pepe, and Fisher (1990) and, apart from a constant term not involving the unknown parameter, is just the likelihood we obtain where the subjects to have been included one by one. There is therefore, operationally, no additional work required to deal with such situations.

The question does arise, however, as to the performance of CRM in such cases. The delayed response can lead to grouping, or we can simply decide upon the grouping by design. Two papers (Goodman et al. 1994; O'Quigley and Shen 1996) have studied the effects of grouping. The more thorough study was that of Goodman et al., in which cohorts of sizes 1, 2, and 3 were evaluated. Broadly speaking the cohort size had little impact upon operating characteristics and the accuracy of final recommendation. O'Quigley and Shen (1996) indicated that for groups of 3, and relatively few patients (16), when the correct level was the highest available level and we started at the lowest or a low level, then we might anticipate some marked drop in performance when contrasted with, say, one by one inclusion. Simple intuition would tell us this, and the differences disappeared for samples of size 25. One by one inclusion tends to maximize efficiency, but should stability throughout the study be an issue, then this extra stability can be obtained through grouping. The cost of this extra stability in terms of efficiency loss appears to be generally small. The findings of Goodman et al. (1994), O'Quigley and Shen (1996), and O'Quigley (1999) contradict the conjecture of Korn et al. (1994) that any grouping would lead to substantial efficiency losses.

2.3.4 ILLUSTRATION

This brief illustration is recalled from O'Quigley and Shen (1996). The study concerned 16 patients. Their toxic responses were simulated from the known dose-toxicity curve. There were six levels in the study; maximum likelihood was used; and the first entered patients were treated at the lowest level. The design was two stage. The true toxic probabilities were $R(d_1) = 0.03$, $R(d_2) = 0.22$, $R(d_3) = 0.45$, $R(d_4) = 0.6$, $R(d_5) = 0.8$, and $R(d_6) = 0.95$. The working model was that given by Equation (2.1) where $\alpha_1 = 0.04$, $\alpha_2 = 0.07$, $\alpha_3 = 0.20$, $\alpha_4 = 0.35$, $\alpha_5 = 0.55$, and $\alpha_6 = 0.70$. The targeted toxicity was given by $\theta = 0.2$ indicating that the best level for the MTD is given by level 2 where the true probability of toxicity is 0.22. A grouped design was used until heterogeneity in toxic responses was observed, patients being included, as for the classical schemes, in groups of 3. The first 3 patients experienced no toxicity at level 1. Escalation then took place to level 2, and the next three patients treated at this level did not experience any toxicity either. Subsequently two out of the three patients treated at level 3 experienced toxicity. Given this heterogeneity in the responses, the maximum likelihood estimator for α now exists and, following a few iterations, could be seen to be equal to

0.715. We then have that $\hat{R}(d_1) = 0.101$, $\hat{R}(d_2) = 0.149$, $\hat{R}(d_3) = 0.316$, $\hat{R}(d_4) = 0.472$, $\hat{R}(d_5) = 0.652$, and $\hat{R}(d_6) = 0.775$. The 10th entered patient is then treated at level 2, for which $\hat{R}(d_2) = 0.149$ because, from the available estimates, this is closest to the target $\theta = 0.2$. The 10th included patient does not suffer toxic effects, and the new maximum likelihood estimator becomes 0.759. Level 2 remains the level with an estimated probability of toxicity closest to the target. This same level is in fact recommended to the remaining patients so that after 16 inclusions the recommended MTD is level 2. The estimated probability of toxicity at this level is 0.212, and a 90% confidence interval for this probability is estimated as (0.07, 0.39).

2.4 STATISTICAL PROPERTIES

Recall that CRM is a class of methods rather than a single method, the members of the class depending on arbitrary quantities chosen by the investigator, such as the form of the model, the spacing between the doses, the starting dose, single or grouped inclusions, the initial dose escalation scheme in two-stage designs, or the prior density chosen for Bayesian formulations. The statistical properties described in this section apply broadly to all members of the class, the members nonetheless maintaining some of their own particularities.

2.4.1 CONVERGENCE

Convergence arguments obtain from considerations of the likelihood. The same arguments apply to Bayesian estimation as long as the prior is other than degenerate, i.e., all the probability mass is not put on a single point. Usual likelihood arguments break down because our models are misspecified. The maximum likelihood estimate, $\hat{R}(d_i) = \psi(d_i, \hat{a}_j)$, exists as soon as we have some heterogeneity in the responses (Silvapulle 1981). We need to assume that the dose toxicity function, $\psi(x, a)$, satisfies the conditions described in Shen and O'Quigley (1996); in particular the condition that, for $i = 1, \dots, k$, there exists a unique a_i such that $\psi(d_i, a_i) = R(d_i)$. Note that the a_i s depend on the actual probabilities of toxicity and are therefore unknown. We also require:

1. For each $0 < t < 1$ and each x , the function

$$s(t, x, a) := t \frac{\psi'}{\psi}(x, a) + (1 - t) \frac{-\psi'}{1 - \psi}(x, a)$$

is continuous and is strictly monotone in a .

2. The parameter a belongs to a finite interval $[A, B]$.

The first condition is standard for estimating equations to have unique solutions. The second imposes no real practical restriction. We will also require the true unknown dose toxicity function, $R(x)$, to satisfy the following conditions:

1. The probabilities of toxicity at $d_1, \dots, d_k$ satisfy: $0 < R(d_1) < \dots < R(d_k) < 1$.

2. The target dose level is $x_0 \in (d_1, \dots, d_k)$ where $|\hat{R}(x_0) - \theta| < |\hat{R}(d_i) - \theta|$, ($i = 1, \dots, k$; $x_0 \neq d_i$).
3. Before writing down the third condition, note that because our model is misspecified it will generally not be true that $\psi(d_i, a_0) \equiv R(d_i)$ for $i = 1, \dots, k$. We will nonetheless require that the working model be not "too distant" from the true underlying dose toxicity curve, and this can be made precise with the help of the set

$$S(a_0) = \{a : |\psi(x_0, a) - \theta| < |\psi(x_i, a) - \theta|, \text{ for all } d_i \neq x_0\} \quad (2.4)$$

The condition we require is that for $i = 1, \dots, k$, $a_i \in S(a_0)$.

Under these conditions, Shen and O'Quigley (1996) showed that the estimator $\hat{a}_n$ converges to a_0 and that the asymptotic distribution of $\sqrt{n}(\hat{a}_n - a_0)$ is $N(0, \sigma^2)$, with $\sigma^2 = \{\psi'(x_0, a_0)\}^{-2}\theta_0(1 - \theta_0)$.

2.4.2 EFFICIENCY

We can use $\hat{\theta}_n = \psi(x_{n+1}, \hat{a}_n)$ to estimate the probability of toxicity at the recommended level x_{n+1} , where $\hat{a}_n$ is the maximum likelihood estimate (O'Quigley 1992). An application of the δ -method shows that the asymptotic distribution of $\sqrt{n}\{\hat{\theta}_n - R(x_0)\}$ is $N\{0, \theta_0(1 - \theta_0)\}$. The estimate then provided by CRM is fully efficient. This is what our intuition would suggest given the convergence properties of CRM and the variance of a binomial variate. What actually takes place in finite samples needs to be investigated on a case by case basis. Nonetheless the relatively broad range of cases studied by O'Quigley (1992) show a mean squared error for the estimated probability of toxicity at the recommended level under CRM to correspond well with the theoretical variance for samples of size n , were all subjects to be experimented at the correct level. Some of the cases studied showed evidence of super-efficiency, translating nonnegligible bias that happens to be in the right direction, while a few others indicated efficiency losses large enough to suggest the potential for improvement.

Large sample results are helpful inasmuch as they provide some assurance as to the basic statistical soundness of the approach. For instance, some suggested approaches using richer parametric models turn out to be not only inefficient but also inconsistent. However, in practice, we are typically interested in behavior at small sample sizes. For some arbitrarily chosen true dose-toxicity relations, authors have studied the relative behavior of competing schemes case by case in terms of percentage of final recommendation, in-trial allocation, probability of toxicity, average number of subjects included, etc. The operating characteristics are to some extent dependent upon the true dose-toxicity relation (Gooley et al. 1994). The choice then of such relations and their influence on the evaluation of the performance of the method under study raises questions of objectivity.

Nonetheless, it can be seen (O'Quigley, Paoletti, and Maccario 2002) that there does exist an optimal scheme making the fullest use of all the information in the experiment. The scheme is optimal in that the mean squared error of estimate of

the probability of toxicity at the recommended dose is less than or equal to all other asymptotically unbiased estimates. Of course, Bayesian schemes could outperform the optimal design by including accurate information. It is also helpful to think in terms of the complementary idea, suboptimality, because we will see that suboptimality can be seen to be equivalent to the concept of incomplete information.

Most experiments have incomplete information in that it is not usually possible to replicate experiments for each subject at each level. Were it possible to experiment each subject independently at each level, then such a scheme would in fact be equivalent to the nonparametric optimal method. In a real experiment each patient provides partial or incomplete information. The monotonicity of toxicity assumption implies that if a subject had a toxic reaction at level d_k ($k \leq 6$) then he or she would necessarily have a toxic reaction at d_ℓ ($k \leq \ell \leq 6$). As for his or her response at levels below d_k we have no information on this. For instance, a subject experiencing a toxicity at d_5 provides the following information:

Dose	d_1	d_2	d_3	d_4	d_5	d_6
Y_k	*	*	*	*	1	1

where an asterisk (*) indicates missing or incomplete information. Should the subject tolerate the treatment at level d_k ($1 \leq k \leq 6$), then he or she would necessarily tolerate the treatment at all levels d_ℓ ($1 \leq \ell \leq k$). Thus, if a subject had been included at d_3 without a toxic response, the experiment would provide the tabulated information:

Dose	d_1	d_2	d_3	d_4	d_5	d_6
Y_k	0	0	0	*	*	*

The above considerations help us understand in a precise way why we describe our data as being incomplete. If all the information were available for each patient, we would know the response at every dose level. In other words, the highest tolerated dose level would be known. For instance, instead of the above two tables, we could imagine a table for a subject for whom DLT appears from dose level 3 (Table 2.2). Of course in a real trial such information is not available. However, in the framework of simulations or probabilistic calculation, complete information can be obtained. As an illustration we take the dose-toxicity relation from O'Quigley, Paoletti, and Maccario (2002). Keeping the dose-toxicity relation from our previous example, we carried out 5000 simulations of the two procedures. Table 2.3 gives the recommendation distribution when the target is 0.2. We denote by $q_k(16)$ the proportion of times that the optimal method recommends level k based on 16 patients and $p_k(16)$ the analogous quantity for the CRM.

TABLE 2.2
Complete Information for a Given Subject

Dose	d_1	d_2	d_3	d_4	d_5	d_6
Y_k	0	0	1	1	1	1

TABLE 2.3

Compared Frequency of Final Recommendations of the Optimal Method and the CRM for a Simulated Example

d_k	1	2	3	4	5	6
R_k	0.05	0.11	0.22	0.35	0.45	0.60
$p_k(16)$	0.05	0.26	0.42	0.21	0.06	0.0
$q_k(16)$	0.04	0.27	0.48	0.17	0.04	0.0

2.4.3 SAFETY

In any discussion on a phase I design, the word *safety* will arise. This translates a central ethical concern. A belief that CRM would tend to treat the early included patients in a study at high dose levels convinced many investigators that without some modification CRM was not safe.

Safety is in fact a statistical property of any method. When faced with some potential realities or classes of realities, we can ask ourselves questions such as what is the probability of toxicity for a randomly chosen patient that has been included in the study or, say, what is the probability of toxicity for those patients entered into the study at the very beginning.

Once we know the realities or classes of realities we are facing, the operating rules of the method, which are obvious and transparent for up- and -down schemes and less transparent for model based schemes such as CRM, then in principle we can calculate the probabilities mentioned above. In practice, these calculations are involved, and we may simply prefer to estimate them to any desired degree of accuracy via simulation.

Theoretical work as well as extensive simulations (O'Quigley, Pepe, and Fisher 1990; O'Quigley and Chevret 1994; O'Quigley and Shen 1996; Ahn 1998; O'Quigley 1999; and Reiner et al. 1998) indicate CRM to be a safer design than any of the commonly used up- and -down schemes in that, for targets of less than $\theta = 0.30$, the probability that a randomly chosen patient suffers a toxicity is lower. Furthermore the probability of being treated at levels higher than the MTD was, in all the studied cases, higher with the standard designs than with CRM.

Were the definition of safety to be widened to include the concept of treating patients at unacceptably low levels, i.e., levels at which the probability of toxicity is deemed too close to zero, then CRM does very much better than the standard designs. This finding is logical given that the purpose of CRM is to concentrate as much experimentation as possible around the prespecified target. In addition it ought be emphasized that we can adjust the CRM to make it as safe as we require by changing the target level. For instance, were we to decrease the target from 0.20 to 0.10, the observed number of toxicities will, on average, be reduced and, in many cases, halved. This is an important point because it highlights the main advantages of the CRM over the standard designs in terms of flexibility and the ability to be adapted to potentially different situations. An alternative way to enhance conservatism is, rather than to choose the closest available dose to the target, systematically to take the dose immediately lower than the target or change the distance measure

used to select the next level to recommend. Safety ought be improved, although the impact such an approach might have on the reliability of final estimation remains to be studied. Some study on this idea has been carried out by Babb et al. (1998).

2.5 MORE COMPLEX CRM DESIGNS

The different up- and -down designs amount to a collection of ad hoc rules for making decisions when faced with accumulating observations. The CRM leans on a model that, although not providing a broad summary of the true underlying probabilistic phenomenon, in view of its being under parameterized, does nonetheless provide structure enabling better control in an experimental situation. In principle at least, a model enables us to go further and accommodate greater complexity. Care is needed, but with some skill in model construction, we may hope to capture some other effects that are necessarily ignored by the rough- and -ready up- and -down designs. The following sections consider some examples.

2.5.1 INCLUSION OF A STOPPING RULE

The usual CRM design requires that a given sample size be determined in advance. However, given the convergence properties of CRM, it may occur in practice that we appear to have settled on a level before having included the full sample size n of anticipated patients. In such a case we may wish to bring the study to an early close, thereby enabling the phase II study to be undertaken more quickly.

One possible approach suggested by O'Quigley, Pepe, and Fisher (1990) would be to use the Bayesian intervals, (ψ_j^-, ψ_j^+) , for the probability of toxicity, $\psi(x_{j+1}, \hat{a}_j)$, at the currently recommended level and when this interval falls within some pre-specified range then stop the study. Another approach would be to stop after some fixed number of subjects have been treated at the same level. Such designs were used by Goodman et al. (1995) and Korn et al. (1994) and have the advantage of great simplicity. The properties of such rules remain to be studied, and we do not recommend their use at this point.

One stopping rule that has been studied in detail (O'Quigley and Reiner 1998) is described here. The idea is based on the convergence of CRM and the idea that as we reach a plateau the accumulating information can enable us to quantify this notion. Specifically, given Ω_j we would like to be able to say something about the levels at which the remaining patients, $j + 1$ to n , are likely to be treated. The quantity we are interested in is

$$\mathcal{P}_{j,n} = \Pr\{x_{j+1} = x_{j+2} = \cdots = x_{n+1} | \Omega_j\}$$

In words, $\mathcal{P}_{j,n}$ is the probability that x_{j+1} is the dose recommended to all remaining patients in the trial as well as being the final recommended dose. Thus, to find $\mathcal{P}_{j,n}$ one needs to determine all the possible outcomes of the trial based on the results known for the first j patients.

The following algorithm achieves the desired result.

1. Construct a complete binary tree on $2^{n-j+1} - 1$ nodes corresponding to all possible future outcomes $(y_{j+1}, \dots, y_n)$. The root is labeled with x_{j+1} .
2. Assuming that $y_{j+1} = 1$, compute the value of $\hat{a}_{j+1}$. Determine the dose level that would be recommended to patient $j + 1$ in this case. Label the left child of the root with this dose level.
3. Repeat step 2, this time with $y_{j+1} = 1$. Determine the dose level that would be recommended to patient $j + 1$. Label the right child of the root with this dose level.
4. Label the remaining nodes of the tree level by level, as in the above steps.

We use the notation $T_{j,n}$ to refer to the tree constructed with this algorithm. Each path in $T_{j,n}$ that starts at the root and ends at a leaf whose nodes all have the same label represents a trial where the recommended dose is unchanged between the $(j + 1)$ st and the $(n + 1)$ st patient. The probability of each such path is given by

$$\{R(x_{j+1})\}^r \{1 - R(x_{j+1})\}^{n-j-r}$$

where r is the number of toxicities along the path. By exclusivity we can sum the probabilities of all such paths to obtain $\mathcal{P}_{j,n}$.

Using $\psi\{x_{j+1}, \hat{a}_j\}$, the current estimate of the toxicity of x_{j+1} , we may estimate the probability of each path by

$$[\psi(x_{j+1}, \hat{a}_j)]^r [1 - \psi(x_{j+1}, \hat{a}_j)]^{n-j-r}$$

Adding up these path estimates yields an estimate of $\mathcal{P}_{j,n}$. Details are given in O'Quigley and Reiner (1998).

2.5.2 PATIENT HETEROGENEITY

As in other types of clinical trials, we are essentially looking for an average effect. Patients of course differ in the way they may react to a treatment, and although hampered by small samples, we may sometimes be in a position to specifically address the issue of patient heterogeneity. One example occurs in patients with acute leukemia where it has been observed that children will better tolerate more aggressive doses (standardized by their weight) than adults. Likewise, heavily pretreated patients are more likely to suffer from toxic side effects than lightly pretreated patients. In such situations we may wish to carry out separate trials for the different groups in order to identify the appropriate MTD for each group. Otherwise we run the risk of recommending an "average" compromise dose level that is too toxic for a part of the population and suboptimal for the other. Usually, clinicians carry out two separate trials or split a trial into two arms after encountering the first DLTs when it is believed that there are two distinct prognostic groups. This has the disadvantage of failing to utilize information common to both groups. A two-sample CRM has been developed so that only one trial is carried out based on information

from both groups. A multisample CRM is a direct generalization, although we must remain realistic in terms of what is achievable in the light of the available sample sizes.

Let I , taking value 1 or 2, be the indicator variable for the two groups. Otherwise, we use the same notation as previously defined. For clarity, we suppose that the targeted probability is the same in both groups and is denoted by θ , although this assumption is not essential to our conclusions.

The dose-toxicity model is now the following:

$$\Pr(Y = 1 \mid X = x, I = 1) = \psi_1(x, a)$$

$$\Pr(Y = 1 \mid X = x, I = 2) = \psi_2(x, a, b)$$

Parameter b measures to some extent the difference between the groups. The functions ψ_1 and ψ_2 are selected in such a way that for each $\theta \in (0, 1)$ and each dose level x there exists (a_0, b_0) satisfying $\psi_1(x, a_0) = \theta$ and $\psi_2(x, a_0, b_0) = \theta$. This condition is satisfied by many function-pairs. The following model has performed well in simulations:

$$\psi_1(x, a) = \frac{\exp(a + x)}{1 + \exp(a + x)}, \quad \psi_2(x, a, b) = \frac{b \exp(a + x)}{1 + b \exp(a + x)}$$

There are many other possibilities, an obvious generalization of the model of O'Quigley, Pepe, and Fisher (1990), arising from Equation (2.1) in which Equation (2.2) applies to group 1 and

$$\alpha_i = \{\tanh(d_i + b) + 1\}/2 \quad (i = 1, \dots, k) \quad (2.5)$$

to group 2. A nonzero value for b indicates group heterogeneity. Let $z_k = (x_k, y_k, I_k)$, $k = 1, \dots, j$ be the outcomes of the first j patients, where I_k indicates to which group the k -th subject belongs, x_k is the dose-level at which the k -th subject is tested, and y_k indicates whether or not the k -th subject suffered a toxic response. In order to estimate the two parameters, one can use a Bayesian estimate or maximum likelihood estimate as for a traditional CRM design. On the basis of the observations z_k , ($k = 1, \dots, j$) on the first j_1 patients in group 1 and j_2 patients in group 2 ($j_1 + j_2 = j$), We can write the likelihood as

$$\mathcal{L}(a, b) = \prod_{i=1}^{j_1} \psi_1(x_i, a)^{y_i} (1 - \psi_1(x_i, a))^{1-y_i} \times \prod_{i=j_1+1}^j \psi_2(x_i, a, b)^{y_i} (1 - \psi_2(x_i, a, b))^{1-y_i}$$

If we denote by $(\hat{a}_j, \hat{b}_j)$ values of (a, b) maximizing this equation after the inclusion of j patients, then the estimated dose-toxicity relations are $\psi_1(x, \hat{a}_j)$ and $\psi_2(x, \hat{a}_j, \hat{b}_j)$, respectively.

If the $(j + 1)$ th patient belongs to group 1, he or she will be allocated at the dose level that minimizes $|\psi_1(x_{j+1}, \hat{a}) - \theta|$ with $x_{j+1} \in \{d_1, \dots, d_k\}$. If the $(j + 1)$ th patient belongs to group 2, the recommended dose level minimizes $|\psi_2(x_{j+1}, \hat{a}_j, \hat{b}_j) - \theta|$.

The trial is carried out as usual: after each inclusion, our knowledge on the probabilities of toxicity at each dose level for either group is updated via the parameters.

It has been shown that under some conditions the recommendations will converge to the right dose level for both groups as well as the estimate of the true probabilities of toxicity at these two levels. Note that it is neither necessary that the two sample sizes be balanced nor that entry into the study be alternating. The Figure 2.2 gives an illustration of a simulated trial carried out with a 2-parameter model. Implementation was based on likelihood estimation, necessitating nontoxicities and a toxicity in each group before the model could be fully fit. Prior to this occurring, dose level escalation followed an algorithm incorporating grade information paralleling that of Section 2.3.3. The design called for shared initial escalation, i.e., the groups were combined until evidence of heterogeneity began to manifest itself. The first DLT occurred in group 2 for the 5th included patient. At this point, the trial was split into two arms with group 2 recommendation being based on $\mathcal{L}(\hat{a}, 0)$ and $\psi_2(x, \hat{a}, 0)$ and group 1 continuing without a model as in Section 2.3.3. Note that there are many possible variants on this design. The first DLT in group 1 was encountered at dose level 6 and led to recommending a lower level to the next patient to be included. For the remainder of the study, allocation for both groups leaned on the model together with the minimization algorithms described above.

2.5.3 PHARMACOKINETIC STUDIES

Statistical modeling of the clinical situation of phase I dose finding studies, such as takes place with the CRM, is relatively recent. Much more fully studied in the phase I context are pharmacokinetics and pharmacodynamics. Roughly speaking,

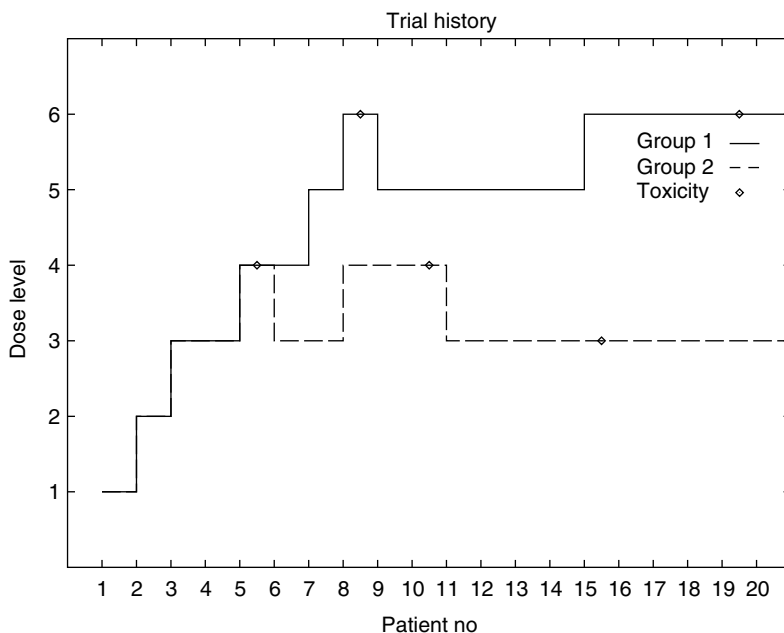


FIGURE 2.2 Simulated trial for two groups.

pharmacokinetics deals with the study of concentration and elimination characteristics of given compounds in specified organ systems, most often blood plasma, whereas pharmacodynamics focuses on how the compounds affect the body. This is a vast subject referred to as PK/PD modeling.

Clearly such information will have a bearing on whether or not a given patient is likely to encounter dose limiting toxicity or, in retrospect, why some patients and not others were able to tolerate some given dose. There are many parameters of interest to the pharmacologist, for example, the area under the concentration time curve, the rate of clearance of the drug, and the peak concentration.

For our purposes, a particular practical difficulty arises in the phase I context, in which any such information becomes available only after the dose has been administered. Most often then the information will be of most use in terms of retrospectively explaining the toxicities. However, it is possible to have pharmacodynamic information and other patient characteristics relating to the patient's ability to synthesize the drugs available before selecting the level at which the patient should be treated.

In principle we can write down any model we care to hypothesize, say one including all the relevant factors believed to influence the probability of encountering toxicity. We can then proceed to estimate the parameters. However, as in the case of patient heterogeneity, we must remain realistic in terms of what can be achieved given the maximum obtainable sample size. Some pioneering work has been carried out here by Piantadosi and Liu (1996), indicating the potential for improved precision by the incorporation of pharmacokinetic information. This is a large field awaiting further exploration.

The strength of CRM is to locate with relatively few patients the target dose level. The remaining patients are then treated at this same level. A recommendation is made for this level. Further studies following the phase I clinical study can now be made, and this is where we see the main advantage of pharmacokinetics. Most patients will have been studied at the recommended level and a smaller amount at adjacent levels. At any of these levels we will have responses and a great deal of pharmacokinetic information. The usual models, in particular the logistic model, can be used to see if this information helps explain the toxicities. If so, we may be encouraged to carry out further studies at higher or lower levels for certain patient profiles indicated by the retrospective analysis to have probabilities of toxicity much lower or much higher than suggested by the average estimate. This can be viewed as the fine tuning and may itself give rise to new, more highly focused phase I studies.

At this point we do not see the utility of a model in which all the different factors are included as regressors. These further analyses are necessarily very delicate requiring great statistical and/or pharmacological skill, and a mechanistic approach based on a catchall model is probably to be advised against.

2.5.4 GRADED TOXICITIES

Although we refer to dose limiting toxicities as a binary (0,1) variable, most studies record information on the degree of toxicity from 0, complete absence of side effects, to 4, life-threatening toxicity. The natural reaction for a statistician is to consider that the response variable, toxicity, has been simplified when going from

5 levels to 2 and that it may help to employ models accommodating multilevel responses.

In fact this is not the way we believe that progress is to be made. The issue is not that of modeling a response (toxicity) at 5 levels but of controlling for dose limiting toxicity, mostly grade 4 but possibly also certain kinds of grade 3. Lower grades are helpful in that their occurrence indicates that we are approaching a zone in which the probability of encountering a dose limiting toxicity is becoming large enough to be of concern. This idea is used implicitly in the two-stage designs described in Section 2.3. If we are to proceed more formally and extract yet more information from the observations, then we need models relating the occurrence of dose limiting toxicities to the occurrence of lower grade toxicities.

In the unrealistic situation in which we can accurately model the ratio of the probabilities of the different types of toxicity, we can make striking gains in efficiency since the more frequently observed lower grade toxicities carry a great deal of information on the potential occurrence of dose limiting toxicities. Such a situation would also allow gains in safety, because at least hypothetically, it would be possible to predict at some level the rate of occurrence of dose limiting toxicities without necessarily having observed very many, the prediction leaning largely upon the model. At the opposite end of the model-hypothesis spectrum we might decide we know nothing about the relative rates of occurrence of the different toxicity types and simply allow the accumulating observations to provide the necessary estimates. In this case it turns out that we neither lose nor gain efficiency, and the method behaves identically to one in which the only information we obtain is whether or not the toxicity is dose limiting. These two situations suggest there may be a middle road using a Bayesian prescription in which very careful modeling can lead to efficiency improvements, if only moderate, without making strong assumptions.

To make this more precise let us consider the case of three toxicity levels, the highest being dose limiting. Let Y_k denote the toxic response at level k , ($k = 1, \dots, 3$). The goal of the trial is still to identify a level of dose whose probability of severe toxicity is closest to a given percentile of the dose-toxicity curve.

A working model for the CRM could be:

$$\begin{aligned}\Pr(Y_k = 3) &= \psi_1(x_k, a) \\ \Pr(Y_k = 2 \text{ or } y_k = 3) &= \psi_2(x_k, a, b) \\ \Pr(Y_k = 1) &= 1 - \psi_2(x_k, a, b)\end{aligned}$$

The contributions to the likelihood are $1 - \psi_2(x_k, a, b)$ when $Y_k = 1$, $\psi_1(x_k, a)$ when $y_k = 3$, and $\psi_2(x_k, a, b) - \psi_1(x_k, a)$ when $Y_k = 2$.

With no prior information, and being able to maximize the likelihood, we obtain exactly the same results as with the more usual one parameter CRM. This is due to the parameter orthogonality. There is therefore no efficiency gain, although there is the advantage of learning about the relationship between the different toxicity types.

Let us imagine that the parameter b is known precisely. The model need not be correctly specified, although b should maintain interpretation outside the model, for instance some simple function of the ratio of grade 3 to grade 2 toxicities.

Efficiency gains are then quite substantial (Paoletti and O'Quigley 1998). This is not of direct practical interest because the assumption of no error in b is completely inflexible; i.e., should we be wrong in our assumed value then this induces a non-correctable bias. In order to overcome this, we have investigated a Bayesian setup in which we use prior information to provide a "point estimate" for b but having uncertainty associated with it, expressed via a prior distribution. Errors can then be overwritten by the data. This work is incomplete at the time of writing, but early results are encouraging.

2.6 COMBINED TOXICITY-EFFICACY STUDIES

O'Quigley, Hughes, and Fenton (2002) presented a class of simple designs to be used for early dose finding studies in HIV. Unlike the classic phase I designs in cancer, these designs have a lot of the phase II flavor about them. Information on efficacy is obtained during the trial and is as important as that relating to toxicity. The designs proposed by O'Quigley, Hughes, and Fenton (2001) incorporate information obtained on viral reduction. The ideas extend immediately to the cancer setting, in which instead of viral reduction we have some other objective measure of response. Initial doses are given from some fixed range of dose regimens. The doses are ordered in terms of their toxic potential. At any dose, a patient can have one of three outcomes; toxicity (whether or not the treatment is otherwise effective), nontoxicity together with an insufficient response, or nontoxicity in the presence of an adequate response. The goal of the study is the identification of the dose leading to the greatest percentage of successes.

One simple approach with encouraging results was the following. A CRM design is used to target some low toxicity level. Information is simultaneously accrued on efficacy. Whenever efficacy is deemed too low at the target toxicity level, then that level and all levels lower than it are removed from investigation. The target toxicity is then increased. Whenever efficacy is sufficiently high for the treatment to be considered successful, then the trial is brought to a close. Rules for making decisions on efficacy are based on the sequential probability ratio test. This class of designs has great flexibility. Starting with the inclusion of the first patient, information on the rate of successes among those patients not suffering toxic side effects is gathered. A true rate of success of p_0 or lower is considered unsatisfactory. A rate p_1 or higher of successes is considered a promising treatment. A model free approach would allow us to determine on the basis of empirical observations which of the above rates is the more likely. This is expensive in terms of patient numbers required. A strong modeling approach uses underparameterized models. These create a structure around which we can home in quickly on the best level. This parallels the use of the CRM for the simpler situation of phase I toxicity studies. Failure to distinguish between the above two hypotheses leads to further experimentation. The experimentation is carried out at a level guided by the toxicity criterion. A conclusion in favor of a success rate greater than p_1 at some level brings the trial to a close with a recommendation for that level. A conclusion in favor of a success rate lower than p_0 at some level leads to that level and all lower levels being removed from further

study. At the same time we increase the target acceptable toxicity from θ to $\theta + \Delta\theta$. The trial then continues at those remaining levels.

Specifically, consider a trial in which $j = 1, \dots, n$ patients may be entered and n is the greatest number of patients that we are prepared to enter. As before, Y_j is a binary random variable (0, 1) where 1 denotes a toxic response. Let V_j be a binary random variable (0, 1) where 1 denotes a response to treatment for the j th patient ($j = 1, \dots, n$). The probability of acceptable viral reduction, given no toxicity, at $X_j = x_j$ is given by

$$Q(x_j) = \Pr(V_j = 1 | X_j = x_j, Y_j = 0) \quad (2.6)$$

so that $P(d_i) = Q(d_i)\{1 - R(d_i)\}$ is the probability of success at dose d_i . The goal of the experiment is to identify the dose level ℓ such that $P(d_\ell) > P(d_i)$ for all i not equal to ℓ .

As for toxicity only experiments, we take $R(x_j) = \Pr(Y_j = 1 | X_j = x_j) = \psi(x_j, a)$. We introduce an additional strong modeling assumption that the second modeling assumption concerns the conditional probability of viral success given absence of toxic side effects. We express this as;

$$Q(x_j) = \Pr(V_j = 1 | X_j = x_j, Y_j = 0) = E(V_j | x_j, Y_j = 0) = \phi(x_j, b) \quad (2.7)$$

where $\phi(x_j, b)$ is a one parameter working model. Thus we assume that $Q(x)$ is also monotonic in x . Since it is possible to construct relatively plausible counterexamples to this assumption, in contrast to the monotonicity assumption for $\psi(x)$, we should consider this to be a stronger assumption. Nonetheless such an assumption may often be reasonable, at least as an approximation. Under these models the success rate can be expressed in terms of the parameters a and b as

$$P(d_i) = \phi(d_i, b)\{1 - \psi(d_i, a)\} \quad (2.8)$$

Once we have at least one toxicity and one nontoxicity together with at least one success and one nonsuccess, then we are in a position to estimate the parameters a and b . The next dose to be used is the one that maximizes $P(d_i)$, $i, \dots, k$.

An efficient compromise structure was proposed by O'Quigley, Hughes, and Fenton (2001) and studied further by Ivanova (2003). Initially we would be happy with a dose level if toxicity is low or absent and the success rate is sufficiently high. Should the success rate turn out not to be high enough at a dose with a prespecified toxicity rate, we will within limits be prepared to increase the toxicity rate. We can, of course, only increase the toxicity rate so far because the objective is to obtain a high overall success rate. The overall success rate will eventually decrease as a result of increasing toxicity. However, before this occurs, should the conditional probability of success increase to a sufficiently high degree, then the trade-off may prove worthwhile. Values for the target toxicity level will start out low, for example $\theta = 0.1$, and this value may be increased subsequently. The amount $\Delta\theta$ by which θ increases and the highest value that is allowed are parameters that can be fixed by the investigator. A special case would be $\Delta\theta = 0$.

We proceed in the following way. Firstly, we use the CRM to target some low rate of toxicities. Estimation is based on likelihood. We exploit the accumulating information at visited levels concerning successes and use this to sequentially test two hypotheses. If d_i is the current level, then we test $H_0: P(d_i) < p_0$ against $H_1: P(d_i) > p_1$. Inability to conclude in favor of either hypothesis leads to further experimentation. Conclusion in favor of H_1 leads to the experiment being brought to a close and level d_i recommended for further study. Conclusion in favor of H_0 leads to level d_i and all lower levels being removed from further experimentation. At this point the target toxicity level is changed from θ to $\theta + \Delta\theta$. As mentioned above, we can allow $\Delta\theta$ to depend on i if desired or even allow $\Delta\theta = 0$. If none of the dose levels manages to achieve a satisfactory rate of success before the toxicity rate θ reaches some threshold, then the whole experiment is deemed a failure as far as the new treatment is concerned. These are composite hypotheses, and the technical details of how to go about testing them is described in O'Quigley, Hughes, and Fenton (2001).

2.7 BAYESIAN APPROACHES

The CRM is often referred to as the Bayesian alternative to the classic up- and -down designs used in phase I studies. Hopefully Section 2.3 makes it clear that there is nothing especially Bayesian about the CRM (see also Storer 1998). In O'Quigley, Pepe, and Fisher (1990), for the sake of simplicity, Bayesian estimators and vague priors were proposed. However, there is nothing to prevent us from working with other estimators, in particular the maximum likelihood estimator.

More fully Bayesian approaches, and not simply a Bayesian estimator, have been suggested for use in the context of phase I trial designs. By more fully we mean more in the Bayesian spirit of inference, in which we quantify prior information observed from outside the trial as well as that solicited from clinicians and pharmacologists. Decisions are made more formally using tools from decision theory. Any prior information can subsequently be incorporated via the Bayes formula into a posterior density that also involves the actual current observations. Given the typically small sample sizes often used, a fully Bayesian approach has some appeal in that we would not wish to waste any relevant information at hand. Unlike the setup described by O'Quigley, Pepe, and Fisher (1990) we could also work with informative priors.

Gatsonis and Greenhouse (1992) consider two-parameter probit and logit models for dose response and study the effect of different prior distributions. Whitehead and Williamson (1998) carried out similar studies but with attention focusing on logistic models and beta priors. Whitehead and Williamson (1998) worked with some of the more classical notions from optimal design for choosing the dose levels in a bid to establish whether much is lost by using suboptimal designs. O'Quigley, Pepe, and Fisher (1990) ruled out criteria based on optimal design due to the ethical criterion of the need to attempt to assign the sequentially included patients at the most appropriate level for the patient. This same point was also emphasized by Whitehead and Williamson (1998).

Whitehead and Williamson (1998) suggest that CRM could be viewed as a special case of their designs with their second parameter being assigned a degenerate prior and thereby behaving as a constant. Although in some senses this view is

technically correct, it can be misleading in that for the single sample case two-parameter CRM and one-parameter CRM are fundamentally different. Two-parameter CRM was seen to behave poorly (O'Quigley, Pepe, and Fisher 1990) and is generally inconsistent (Shen and O'Quigley 1996). We have to view the single parameter as necessary in the homogeneous case and not simply a parametric restriction to facilitate numerical integration (see also Section 2.6.1). This was true even when the data were truly generated by the same two-parameter model, an unintuitive finding at first glance but one that makes sense in the light of the comments in 2.6.1.

We do not believe it will be possible to demonstrate large sample consistency for either the Gatsonis and Greenhouse (1992) approach or the Whitehead and Williamson approach as was done for CRM by Shen and O'Quigley (1996). It may well be argued that large sample consistency is not very relevant in such typically small studies. However, it does provide some theoretical comfort and hints that for finite samples things also might work out satisfactorily. At the very least, if we fail to achieve large sample consistency, then we might carry out large numbers of finite sample studies simulating most often under realities well removed from our working model. This was done by O'Quigley, Pepe, and Fisher (1990) for the usual underparameterized CRM. Such comparisons remain to be done for the Bayesian methods discussed here. Nonetheless we can conclude that a judicious choice of model and prior not running into serious conflict with the subsequent observations may help inference in some special cases.

A quite different Bayesian approach has been proposed by Babb, Rogatko, and Zacks (1998). The context is fully Bayesian. Rather than aim to concentrate experimentation at some target level as does CRM, the aim here is to escalate as fast as possible toward the MTD while sequentially safeguarding against overdosing. This is interesting in that it could be argued that the aim of the approach translates in some ways more directly the clinician's objective than does CRM. Model misspecification was not investigated and would be an interesting area for further research. The approach appears promising, and the methodology may be a useful modification of CRM when primary concern is on avoiding overdosing and we are in a position to have a prior on a two parameter function. As above, there may be concerns about large sample consistency when working with a design that tends to settle on some level.

A promising area for Bayesian formulations is one where we have little overall knowledge of any dose-toxicity relationship but we may have some possibly considerable knowledge of some secondary aspect of the problem. Consider the two group case. For the actual dose levels we are looking for we may know almost nothing. Uninformative Bayes or maximum likelihood would then seem appropriate. But we may well know that one of the groups, for example a group weakened by extensive prior therapy, is most likely to have a level strictly less than that for the other group. Careful parameterization would enable this information to be included as a constraint. However, rather than work with a rigid and unmodifiable constraint, a Bayesian approach would allow us to specify the anticipated direction with high probability while enabling the accumulating data to override this assumed direction if the two run into serious conflict. Exactly the same idea could be used in a case where we believe there may be group heterogeneity but it would be very unlikely

the correct MTDs differ by more than a single level. Incorporating such information will improve efficiency. As already mentioned in Section 2.5, under some strong assumptions we can achieve clear efficiency gains when incorporating information on the graded toxicities. Such gains can be wiped out and even become negative through bias when the assumptions are violated. Once again, a Bayesian setup, opens the possibility of compromise so that constraints become modifiable in the light of accumulating data.

REFERENCES

1. Ahn, C. (1998) An evaluation of phase I cancer clinical trial designs. *Stat. in Med.*, 17, 1537–49.
2. Anbar, D. (1977) The application of stochastic methods to the bioassay problem. *J. Statist. Planning Inference*, 1, 191–206.
3. Anbar, D. (1978) A stochastic Newton–Raphson method. *J. Statist. Planning Inference*, 2, 153–63.
4. Anbar, D. (1984) Stochastic approximation methods and their use in bioassay and phase I clinical trials. *Commun. Statist.*, 13, 2451–2467.
5. Babb J., Rogatko A., and Zacks S. (1998) Cancer phase I clinical trials: efficient dose escalation with overdose control. *Stat. in Med.*, 17, 1103–20.
6. Chevret, S. (1993) The continual reassessment method in cancer phase I clinical trials: a simulation study. *Stat. in Med.*, 12, 1093–1108.
7. Faries, D. (1994) Practical modifications of the continual reassessment method for phase I cancer clinical trials. *J. Biopharm. Statist.*, 4, 147–64.
8. Gatsonis, C. and Greenhouse, J.B. (1992) Bayesian methods for phase I clinical trials. *Stat. in Med.*, 11, 1377–1389.
9. Goodman, S., Zahurak, M.L., and Piantadosi, S. (1995) Some practical improvements in the continual reassessment method for phase I studies. *Stat. in Med.*, 14, 1149–61.
10. Gooley, T.A., Martin, P.J., Fisher, L.D., and Pettinger, M. (1994) Simulation as a design tool for phase I/II clinical trials. An example from bone marrow transplantation. *Controlled Clin. Trials*, 15, 450–62.
11. Huber, P.J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. 6th Berkeley Symp.*, 1, 221–233.
12. Ivanova, A. (2003) A new dose-finding design for bivariate outcomes. *Biometrics* 57, 1018–29.
13. Korn, E.L., Midthune, D., Chen, T.T., Rubinstein, L.V., Christian, M.C., and Simon, R. (1994) A comparison of two phase I trial designs. *Stat. in Med.*, 13, 1799–1806.
14. Mick, R. and Ratain, M.J. (1993) Model-guided determination of maximum tolerated dose in phase I clinical trials: evidence for increased precision. *J.Natl. Cancer Inst.*, 85, 217–223.
15. Moller, S. (1995) An extension of the continual reassessment method using a preliminary up and down design in a dose finding study in cancer patients in order to investigate a greater number of dose levels. *Stat. in Med.*, 14, 911–22.
16. Natarajan, L. and O’Quigley J. (2003) Interval estimates of the probability of toxicity at the maximum tolerated dose for small samples. *Stat. in Med.*, 22, 1829–36.
17. O’Quigley, J. (1992) Estimating the probability of toxicity at the recommended dose following a phase I clinical trial in cancer. *Biometrics*, 48, 853–862.

18. O'Quigley, J., Pepe, M., Fisher, L. (1990) Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics*, 46, 33–48.
19. O'Quigley, J. and Chevret, S. (1991) Methods for dose finding studies in cancer clinical trials: a review and results of a Monte Carlo study. *Stat. in Med.*, 10, 1647–1664.
20. O'Quigley, J. and Shen L.Z. (1996) Continual reassessment method: a likelihood approach. *Biometrics*, 52, 163–174.
21. O'Quigley, J. and Reiner E., (1998) A stopping rule for the continual reassessment method. *Biometrika*, 85, 741–48.
22. O'Quigley, J. (1999) Another look at two phase I clinical trial designs (with commentary). *Stat. in Med.*, 18, 2683–92.
23. O'Quigley, J., Shen, L., and Gamst, A. (1999) Two sample continual reassessment method. *J. Biopharm. Statist.*, 9, 17–44.
24. O'Quigley, J., Hughes, M., and Fenton, T. (2001) Dose finding designs for HIV studies. *Biometrics* 57, 1018–29.
25. O'Quigley J., Paoletti, X., and Maccario, J. (2002) Non-parametric optimal design in dose finding studies. *Biostatistics* 3, 51–56.
26. Paoletti, X. and O'Quigley, J. (1998) Using graded toxicities in phase I trial design. *Technical Report 4, Biostatistics Group, University of California at San Diego*.
27. Piantadosi, S. and Liu, G. (1996) Improved designs for dose escalation studies using pharmacokinetic measurements. *Stat. in Med.*, 15, 1605–18.
28. Reiner E., Paoletti, X., and O'Quigley (1998) Operating characteristics of the standard phase I clinical trial design. *Comp. Statist. and Data Anal.*, 30, 303–315.
29. Robbins, H. and Monro, S. (1951) A stochastic approximation method. *Annals of Math. Stat.*, 29, 351–56.
30. Shen, L.Z. and O'Quigley, J. (1996) Consistency of continual reassessment method in dose finding studies. *Biometrika*, 83, 395–406.
31. Simon, R., Freidlin, B., Rubinstein, L., Arbuck, S., Collins, J., and Christian, M. (1997) Accelerated titration designs for phase I clinical trials in oncology. *J. Natl. Cancer Inst.*, 89, 1138–1147.
32. Silvapulle, M.J. (1981) On the existence of maximum likelihood estimators for the binomial response models. *J.R. Stat. Soc. B*, 43, 310–313.
33. Storer, B.E. (1989) Design and analysis of phase I clinical trials. *Biometrics*, 45, 925–937.
34. Storer, B.E. (1993) Small-sample confidence sets for the MTD in a phase I clinical trial. *Biometrics*, 49, 1117–25.
35. Storer, B.E. (1998) Phase I clinical trials. *Encyclopedia of Biostatistics*, Wiley, New York.
36. Whitehead, J. and Williamson, D. (1998) Bayesian decision procedures based on logistic regression models for dose-finding studies. *J. Biopharm. Stat.*, 8, 445–67.
37. Wu, C.F.J. (1985) Efficient sequential designs with binary data, *J. Amer. Stat. Assoc.*, 80, 974–984.
38. Wu, C.F.J. (1986) Maximum likelihood recursion and stochastic approximation in sequential designs, in *Adaptive Statistical Procedures and Related Topics*, ed. J. van Ryzin, Institute of Mathematical Statistics Monograph 8, Hayward, CA: Institute of Mathematical Statistics, 298–313.

3 Choosing a Phase I Design

Barry E. Storer

CONTENTS

3.1	Introduction and Background.....	59
3.2	Designs for Phase I Trials	61
3.2.1	Initial Dose Level and Dose Spacing	61
3.2.2	Traditional Escalation Algorithms	61
3.2.3	A Bayesian Approach: The Continual Reassessment Method.....	63
3.2.4	Storer's Two-Stage Design	64
3.2.5	Other Approaches	66
3.3	Choosing a Phase I Design.....	67
3.3.1	Specifying the Dose–Response Curve	67
3.3.2	Specifying the Designs	68
3.3.2.1	Traditional 3 + 3 Design	68
3.3.2.2	Traditional Best-of-5 Design.....	68
3.3.2.3	Storer's Two-Stage Design	68
3.3.2.4	Non-Bayesian CRM Design	68
3.3.3	Simulation and Results	69
3.4	Summary and Conclusion	72
	References	74

3.1 INTRODUCTION AND BACKGROUND

Although the term phase I is sometimes applied generically to almost any “early” trial, in cancer drug development it usually refers specifically to a dose finding trial whose major endpoint is toxicity. The goal is to find the highest dose of a potential therapeutic agent that has acceptable toxicity; this dose is referred to as the MTD (maximum tolerable dose) and is presumably the dose that will be used in subsequent phase II trials evaluating efficacy. Occasionally, one may encounter trials that are intermediate between phase I and phase II and are referred to as phase IB trials. This is a more heterogeneous group, but typically includes trials that are evaluating some measure of biologic efficacy over a range of doses that have been found to have acceptable toxicity in a phase I trial. In that context the trial determining MTD may be referred to as a phase IA trial. This chapter will focus exclusively on phase I trials with a toxicity endpoint.

What constitutes acceptable toxicity, of course, depends on the potential therapeutic benefit of the drug. There is an implicit assumption with most anticancer

agents that there is a positive correlation between toxicity and efficacy, but most drugs that will be evaluated in phase I trials will prove ineffective at any dose. The problem of defining an acceptably toxic dose is complicated by the fact that patient response is heterogeneous: at a given dose, some patients may experience little or no toxicity, while others may have severe or even fatal toxicity. Since the response of the patient will be unknown before the drug is given, acceptable toxicity is typically defined with respect to the frequency of toxicity in the population as a whole. For example, given a toxicity grading scheme ranging from 0 to 5 (none, mild, moderate, severe, life-threatening, fatal), one might define the MTD as the dose where, on average, one out of three patients would be expected to experience a grade 3 or worse toxicity. In that case, grade 3 or worse toxicity in an individual patient would be referred to as “dose limiting toxicity” (DLT). The definition of DLT may vary from one trial to another depending on what toxicities are expected and how manageable they are.

When defined in terms of the presence or absence of DLT, the MTD can be represented as a quantile of a dose–response curve. Notationally, if Y is a random variable whose possible values are 1 and 0, respectively, depending on whether a patient does or does not experience DLT, and for dose d we have $\psi(d) = \Pr(Y = 1|d)$, then the MTD is defined by $\psi(d_{MTD}) = \theta$, where θ is the desired probability of toxicity. Alternately, one could define Y to be the random variable representing the threshold dose at which a patient would experience DLT. The distribution of Y is referred to as a tolerance distribution, and the dose–response curve is the cumulative distribution function for Y , so that the MTD would be defined by $\Pr(Y \leq d_{MTD}) = \theta$. For a given sample size, the most effective way of estimating this quantile would be from a sample of threshold doses. Such data are nearly impossible to gather, however, as it is impractical to give each patient more than a small number of discrete doses. Further, the data obtained from sequential administration of different doses to the same patient would almost surely be biased, as one could not practicably distinguish the cumulative effects of the different doses from the acute effects of the current dose level. For this reason, almost all phase I trials involve the administration of only a single dose level to each patient and the observation of the frequency of occurrence of DLT in all patients treated at the same dose level.

There are two significant constraints on the design of a phase I trial. The first is the ethical requirement to approach the MTD from below, so that one must start at a dose level believed almost certainly to be below the MTD and gradually escalate upward. The second is the fact that the number of patients typically available for a phase I trial is relatively small, say 15–30, and is not driven traditionally by rigorous statistical considerations requiring a specified degree of precision in the estimate of MTD. The pressure to use only small numbers of patients is large—literally dozens of drugs per year may come forward for evaluation, and each combination with other drugs, each schedule, and each route of administration requires a separate trial. Furthermore, the number of patients for whom it is considered ethically justified to participate in a trial with little evidence of efficacy is limited. The latter limitation also has implications for the relevance of the MTD in subsequent phase II trials of efficacy. Since the patient populations are different, it is not clear that the MTD estimated in one population will yield the same result when implemented in another.

3.2 DESIGNS FOR PHASE I TRIALS

As a consequence of the above considerations, the traditional phase I trial design utilizes a set of fixed dose levels that have been specified in advance, that is $d \in \{d_1, d_2, \dots, d_K\}$. The choice of the initial dose level d_1 and the dose spacing are discussed in more detail below. Beginning at the first dose level, small numbers of patients are entered, typically three to six, and the decision to escalate or not depends on a prespecified algorithm related to the occurrence of DLT. When a dose level is reached with unacceptable toxicity, then the trial is stopped.

3.2.1 INITIAL DOSE LEVEL AND DOSE SPACING

The initial dose level is generally derived either from animal experiments, if the agent in question is completely novel; or by conservative consideration of previous human experience, if the agent in question has been used before but with a different schedule, route of administration, or with other concomitant drugs. A common starting point based on the former is from 1/10 to 1/3 of the mouse LD_{10} , the dose that kills 10% of mice, adjusted for the size of the animal on a per kilogram basis or by some other method.

Subsequent dose levels are determined by increasing the preceding dose level by decreasing multiples, a typical sequence being $\{d_1, d_2 = 2d_1, d_3 = 1.67d_2, d_4 = 1.5d_3, d_5 = 1.4d_4, \text{ and thereafter } d_{k+1} = 1.33d_k, \dots\}$. Such sequences are often referred to as “modified Fibonacci”, although in a true Fibonacci sequence the increments would be approximately 2, 1.5, 1.67, 1.60, 1.63, and then 1.62 thereafter, converging on the golden ratio. Note that after the first few increments the dose levels are equally spaced on a log scale. With some agents, particularly biological agents, the dose levels may be determined by log, i.e., $\{d_1, d_2 = 10d_1, d_3 = 10d_2, d_4 = 10d_3, \dots\}$, or approximate half-log, i.e., $\{d_1, d_2 = 3d_1, d_3 = 10d_1, d_4 = 10d_2, d_5 = 10d_3, \dots\}$, spacing.

3.2.2 TRADITIONAL ESCALATION ALGORITHMS

A wide variety of dose escalation rules may be used. For purposes of illustration, we describe the following, which is often referred to as the traditional “3 + 3” design. Beginning at $k = 1$:

- [A] Evaluate 3 patients at d_k .
 - [A1] If 0 of 3 have DLT, then increase dose to d_{k+1} and go to [A].
 - [A2] If 1 of 3 have DLT, then go to [B].
 - [A3] If ≥ 2 of 3 have DLT, then go to [C].
- [B] Evaluate an additional 3 patients at d_k .
 - [B1] If 1 of 6 have DLT, then increase dose to d_{k+1} and go to [A].
 - [B2] If ≥ 2 of 6 have DLT, then go to [C].
- [C] Discontinue dose escalation.

If the trial is stopped, then the dose level below that at which excessive DLT was observed is the MTD. Some protocols may specify that if only 3 patients were

evaluated at that dose level then an additional 3 should be entered for a total of 6, and that process should proceed downward, if necessary, so that the MTD becomes the highest dose level where no more than 1 toxicity is observed in 6 patients. The actual θ that is desired is generally not defined when such algorithms are used but implicitly $0.17 \leq \theta \leq 0.33$, so we could take $\theta \approx 0.25$.

Another example of a dose escalation algorithm, referred to as the “best-of-5” design, is described here. Again, the value of θ is not explicitly defined, but one could take $\theta \approx 0.40$. Beginning at $k = 1$:

- [A] Evaluate 3 patients at d_k .
 - [A1] If 0 of 3 have DLT, then increase dose to d_{k+1} and go to [A].
 - [A2] If 1 or 2 of 3 have DLT, then go to [B].
 - [A3] If 3 of 3 have DLT, then go to [D].
- [B] Evaluate an additional 1 patient at d_k .
 - [B1] If 1 of 4 have DLT, then increase dose to d_{k+1} and go to [A].
 - [B2] If 2 of 4 have DLT, then go to [C].
 - [B3] If 3 of 4 have DLT, then go to [D].
- [C] Evaluate an additional 1 patient at d_k .
 - [C1] If 2 of 5 have DLT, then increase dose to d_{k+1} and go to [A].
 - [C3] If 3 of 5 have DLT, then go to [D].
- [D] Discontinue dose escalation.

Although traditional designs reflect an empirical common sense approach to the problem of estimating the MTD under the noted constraints, only brief reflection is needed to see that the determination of MTD will have a rather tenuous statistical basis. Consider the outcome of a trial employing the “3 + 3” design where the frequency of DLT for dose levels d_1 , d_2 , and d_3 is 0 of 3, 1 of 6, and 2 of 6, respectively. Ignoring the sequential nature of the escalation procedure, the pointwise 80% confidence intervals for the rate of DLT at the three dose levels are, respectively, 0–0.54, 0.02–0.51, and 0.09–0.67. Although the middle dose would be taken as the estimated MTD, there is not even reasonably precise evidence that the toxicity rate for *any* of the three doses is either above or below the implied θ of approximately 0.25.

Crude comparisons among traditional dose escalation algorithms can be made by examining the level-wise operating characteristics of the design, i.e., the probability of escalating to the next dose level given an assumption regarding the underlying probability of DLT at the current dose level. Usually this calculation is a function of simple binomial success probabilities. For example, in the “3 + 3” algorithm described above, the probability of escalation is $B(0, 3; \psi(d)) + B(1, 3; \psi(d))$ $B(0, 3; \psi(d))$, where $B(r, n; \psi(d))$ is the probability of r successes (toxicities) out of n trials (patients) with underlying success probability at the current dose level $\psi(d)$. The probability of escalation can then be plotted over a range of $\psi(d)$, as is done in [Figure 3.1](#) for the two algorithms described above. Although it is obvious from such a display that one algorithm is considerably more aggressive than another, the level-wise operating characteristics do not provide much useful insight into whether or not a particular design will tend to select an MTD that is close to the target. More useful

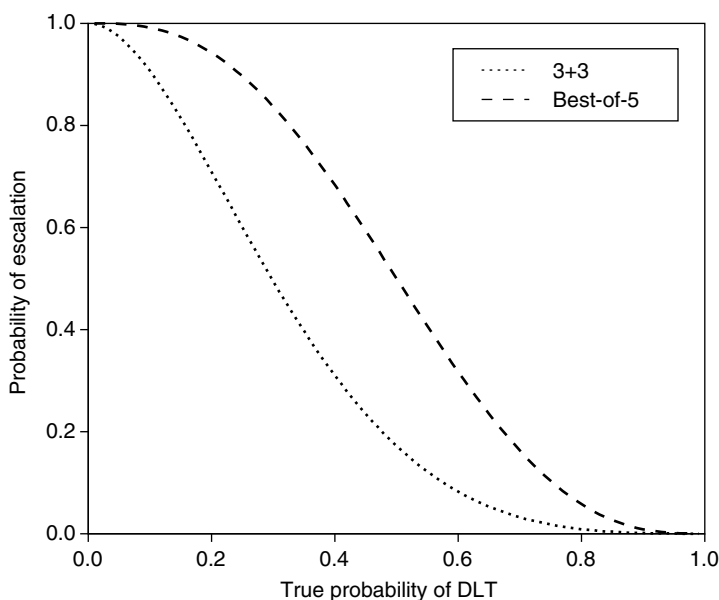


FIGURE 3.1 Levelwise operating characteristics of two traditional dose escalation algorithms. The probability of escalating to the next higher dose level is plotted as a function of the true probability of DLT at the current dose.

approaches to choosing among traditional designs and the other designs described below are discussed in Section 3.3.

3.2.3 A BAYESIAN APPROACH: THE CONTINUAL REASSESSMENT METHOD

The small sample size and low information content in the data derived from traditional methods have suggested to some the usefulness of Bayesian methods to estimate the MTD. In principle, this approach allows one to combine any prior information available regarding the value of the MTD with subsequent data collected in the phase I trial to obtain an updated estimate reflecting both.

The most clearly developed Bayesian approach to phase I design is the continual reassessment method (CRM) proposed by O'Quigley and colleagues.^{1,2} From among a small set of possible dose levels, say $\{d_1, \dots, d_6\}$, experimentation begins at the dose level that the investigators believe, based on all available information, is the most likely to have an associated probability of DLT equal to the desired θ . It is assumed that there is a simple family of monotone dose-response functions ψ such that for any dose d and probability of toxicity p there exists a unique a where $\psi(d, a) = p$ and in particular that $\psi(d_{MTD}, a_0) = \theta$. An example of such a function is $\psi(d, a) = [(\tanh d + 1)/2]^a$. Note that ψ is not assumed to be necessarily a dose-response function relating a characteristic of the dose levels to the probability of toxicity. That is, d does not need to correspond literally to the dose of a drug. In fact, the treatments at each of the dose levels may be completely unrelated as long

as the probability of toxicity increases from each dose level to the next; in this case d could be just the index of the dose levels. The uniqueness constraint implies in general the use of one-parameter models and explicitly eliminates popular two-parameter dose-response models like the logistic. In practice, the latter have a tendency to become “stuck” and oscillate between dose levels when any data configuration leads to a large estimate for the slope parameter.

A prior distribution $g(a)$ is assumed for the parameter a such that for the initial dose level, for example d_3 , either $\int_0^\infty \psi(d_3, a)g(a)da = \theta$ or, alternatively, $\psi(d_3, \mu_a) = \theta$, where $\mu_a = \int_0^\infty ag(a)da$. The particular prior used should also reflect the degree of uncertainty present regarding the probability of toxicity at the starting dose level; in general this will be quite vague.

After each patient is treated and the presence or absence of toxicity observed, the current distribution $g(a)$ is updated along with the estimated probabilities of toxicity at each dose level, calculated by either of the methods above.¹ The next patient is then treated at the dose level minimizing some measure of the distance between the current estimate of the probability of toxicity and θ . After a fixed number n of patients has been entered sequentially in this fashion, the dose level selected as the MTD is the one that would be chosen for a hypothetical $n + 1$ st patient.

An advantage of the CRM design is that it makes full use of all the data at hand in order to choose the next dose level. Even if the dose-response model used in updating is misspecified, CRM will tend eventually to select the dose level that has a probability of toxicity closest to θ ,³ although its practical performance should be evaluated in the small sample setting typical of phase I trials. A further advantage is that, unlike traditional algorithms, the design is easily adapted to different values of θ .

In spite of these advantages, some practitioners object philosophically to the Bayesian approach, and it is clear in the phase I setting that the choice of prior can have a measurable effect on the estimate of MTD.⁴ However, the basic framework of CRM can easily be adapted to a non-Bayesian setting and can conform in practice more closely to traditional methods.⁵ For example, there is nothing in the approach that prohibits one from starting at the same low initial dose as would be common in traditional trials or from updating after groups of three patients rather than single patients. In fact, the Bayesian prior can be abandoned entirely, and the updating after each patient can be fully likelihood based. Without a prior, however, the dose-response model cannot be fit to the data until there is some heterogeneity in outcome, i.e., at least one patient with DLT and one patient without DLT. Thus, some simple rules are needed to guide the dose escalation until heterogeneity is achieved. A number of other modifications to the CRM design have been proposed that address practical issues of implementation. These include limiting the rate of dose escalation,⁶ stopping rules based on the width of the posterior probability interval,⁷ and interpolation between doses.⁸

3.2.4 STORER'S TWO-STAGE DESIGN

Storer^{9,10} has explored a combination of more traditional methods implemented in such a way as to minimize the numbers of patients treated at low dose levels and to

focus sampling around the MTD; these methods also utilize an explicit dose–response framework to estimate the MTD.

The design has two stages and uses a combination of simple dose-escalation algorithms. The first stage assigns single patients at each dose level and escalates upward until a patient has DLT or downward until a patient does not have DLT. Algorithmically, beginning at $k = 1$:

- [A] Evaluate one patient at d_k .
 - [A1] If no patient has had DLT, then increase dose to d_{k+1} and go to [A].
 - [A2] If all patients have had DLT, then decrease dose to d_{k-1} and go to [A].
 - [A3] If at least one patient has and has not had DLT, then if the current patient has not had DLT, go to [B]; otherwise decrease the dose to d_{k-1} and go to [B].

Note that the first stage meets the requirement for heterogeneity in response needed to start off a likelihood-based CRM design and could be used for that purpose. The second stage incorporates a fixed number of cohorts of patients. If $\theta = 1/3$, then it is natural to use cohorts of size three, as follows:

- [B] Evaluate three patients at d_k .
 - [B1] If 0 of 3 have DLT, then increase dose to d_{k+1} and go to [B].
 - [B2] If 1 of 3 have DLT, then go to [B].
 - [B3] If ≥ 2 of 3 have DLT, then decrease dose to d_{k-1} and go to [B].

After completion of the second stage, a dose–response model is fit to the data and the MTD estimated by maximum likelihood or other method. For example, one could use a logistic model where $\text{logit}(\psi(d)) = \alpha + \beta \log(d)$, whence the estimated MTD is $\log(d_{\text{MTD}}) = (\text{logit}(\theta) - \hat{\alpha})/\hat{\beta}$. A two-parameter model is used here in order to make fullest use of the final sample of data; however, as noted above, two-parameter models have undesirable properties for purposes of dose escalation. In order to obtain a meaningful estimate of the MTD, one must have $0 < \hat{\beta} < \infty$. If this is not the case, then one needs either to add additional cohorts of patients or substitute a more empirical estimate, such as the last dose level or hypothetical next dose level.

As noted, the algorithm described above is designed with a target $\theta = 1/3$ in mind. Although other quantiles could be estimated from the same estimated dose–response curve, a target θ different from $1/3$ would probably lead one to use a modified second-stage algorithm.

Extensive simulation experiments using this trial design in comparison to more traditional designs demonstrated the possibility of reducing the variability of point estimates of the MTD and reducing the proportion of patients treated at very low dose levels without markedly increasing the proportion of patients treated at dose levels where the probability of DLT is excessive. Storer¹⁰ also evaluated different methods of providing confidence intervals for the MTD and found that standard likelihood-based methods that ignore the sequential sampling scheme are often markedly anti-conservative; these methods included the delta method, a method based on Fieller's theorem, and a likelihood ratio method. More accurate confidence sets can be

constructed by simulating the distribution of any of those test statistics at trial values of the MTD; however, the resulting confidence intervals are often extremely wide. Furthermore, the methodology is purely frequentist and may be unable to account for minor variations in the implementation of the design when a trial is conducted.

With some practical modifications, the two-stage design described above has been implemented in a real phase I trial.¹¹ The major modifications included: (a) a provision to add additional cohorts of three patients if necessary until the estimate of β in the fitted logistic model becomes positive and finite; (b) a provision that if the estimated MTD is higher than the highest dose level at which patients have actually been treated, the latter will be used as the MTD; and (c) a provision to add additional intermediate dose levels if in the judgment of the protocol chair the nature or frequency of toxicity at a dose level precludes further patient accrual at that dose level.

3.2.5 OTHER APPROACHES

Although not common in practice, it is useful to consider the case where the major outcome defining toxicity is a continuous measurement, for example the nadir WBC. This may or may not involve a fundamentally different definition of the MTD in terms of the occurrence of DLT. For example, suppose that DLT is determined by the outcome $Y < c$ where c is a constant and we have $Y \sim N(\alpha + \beta d, \sigma^2)$. Then $d_{MTD} = (c - \alpha - \Phi^{-1}(\theta)\sigma)/\beta$ has the traditional definition that the probability of DLT is θ . The use of such a model in studies with small sample size makes some distributional assumption imperative. Some sequential design strategies in this context have been described by Eichhorn et al.¹²

Alternatively, the MTD might be defined in terms of the mean response, i.e., the dose where $E(Y) = c$. For the same simple linear model above, we then have that $d_{MTD} = (c - \alpha)/\beta$. An example of a two-stage design using a regression model for WBC is given in Mick et al.¹³ Fewer distributional assumptions are needed to estimate d_{MTD} , and stochastic approximation techniques might be applied in the design of trials with such an endpoint.¹⁴ Nevertheless, the use of a mean response to define MTD is not generalizable across drugs with different or multiple toxicities and consequently has received little attention in practice.

Another Bayesian approach to phase I design has been described by Babb et al.¹⁵ The approach is referred to as EWOC (escalation with overdose control). The general framework is similar to that of CRM, and the MTD has the usual definition in terms of the probability of DLT; however, in contrast to CRM, the MTD is related explicitly to an underlying continuous tolerance distribution. The dose for each patient is selected such that based on all available information the posterior probability that the dose exceeds the MTD is equal to α . The feasibility bound α controls the aggressiveness of the escalation; a typical value would be $\alpha \approx 0.25$. In the usual case that there are a few fixed dose levels available for testing, additional tolerance parameters are used to select one that is closest to the optimal exact dose. Note that, unlike CRM, the dose chosen during dose escalation is not necessarily the one estimated to be closest to the MTD. After a predetermined number of patients have been evaluated, the final estimate of MTD is determined by minimizing the posterior expected loss with respect to some loss function.

3.3 CHOOSING A PHASE I DESIGN

As noted above, only limited information regarding the suitability of a phase I design can be gained from the levelwise operating characteristics shown in Figure 3.1. Furthermore, for designs like CRM, which depend on data from prior dose levels to determine the next dose level, it is not even possible to specify a levelwise operating characteristic.

Useful evaluations of phase I designs must involve the entire dose–response curve, which of course is unknown. Many simple designs for which the levelwise operating characteristics can be specified can be formulated as discrete Markov chains.⁹ The states in the chain refer to treatment of a patient or group of patients at a dose level, with an absorbing state corresponding to the stopping of the trial. For various assumptions about the true dose–response curve, one can then calculate exactly many quantities of interest, such as the number of patients treated at each dose level, from the appropriate quantities determined from successive powers of the transition probability matrix $\mathbf{P}$. Such calculations are fairly tedious, however, and do not accommodate designs with nonstationary transition probabilities, such as CRM. Nor do they allow one to evaluate any quantity derived from all of the data, such as the MTD estimated after following Storer’s two-stage design.

For these reasons, simulations studies are the only practical tool for evaluating phase I designs. As with exact computations, one needs to specify a range of possible dose–response scenarios and then simulate the outcome of a large number of trials under each scenario. Here we give an example of such a study in order to illustrate the kinds of information that can be used in the evaluation and some of the considerations involved in the design of the study. This study has also been presented in Storer.¹⁶ Other examples of simulation studies comparing phase I designs are Korn et al.¹⁷ and Ahn.¹⁸

3.3.1 SPECIFYING THE DOSE–RESPONSE CURVE

We follow the modified Fibonacci spacing described in Section 3.2. For example, in arbitrary units we have $d_1 = 100.0$, $d_2 = 200.0$, $d_3 = 333.3$, $d_4 = 500.0$, $d_5 = 700.0$, $d_6 = 933.3$, $d_7 = 1244.4$, ... We also define hypothetical dose levels below d_1 that successively halve the dose above, i.e., $d_0 = 50.0$, $d_{-1} = 25.0$, ... The starting dose is always d_1 , and we assume that the true MTD is four dose levels higher at d_5 , with $\theta = 1/3$. In order to define a range of dose–response scenarios, we vary the probability of toxicity at d_1 from 0.01 to 0.20 in increments of 0.01 and graph our results as a function of that probability. The true dose–response curve is determined by assuming that a logistic model holds on the log scale. In the usual formulation one would have $\text{logit}(\psi(d)) = \alpha + \beta \log d$. In the present setup, we specify d_1 , $\psi(d_1)$, and that $\psi(d_5) = 1/3$, whence $\beta = (\text{logit}(1/3) - \text{logit}(\psi(d_1)))/\Delta$, where $\Delta = \log d_5 - \log d_1$, and $\alpha = \text{logit}(\psi(d_1)) - \beta \log d_1$.

Varying the probability of DLT at d_1 while holding the probability at d_5 fixed at θ results in a sequence of dose–response curves ranging from relatively steep to relatively flat. An even greater range could be encompassed by also varying the number of dose levels between the starting dose and the true MTD, which of course

need not be exactly at one of the predetermined dose levels. The point is to study the sensitivity of the designs to features of the underlying dose–response curve, which obviously is unknown.

3.3.2 SPECIFYING THE DESIGNS

This simulation will evaluate the two traditional designs described above, Storer’s two-stage design, and a non-Bayesian CRM design. It is important to make the simulation as realistic as possible in terms of how an actual clinical protocol would be implemented, or at least to recognize what differences might exist. For example, the simulation does not place a practical limit on the highest dose level, although it is rare for any design to escalate beyond d_{10} . An actual protocol might have an upper limit on the number of dose levels with a provision for how to define the MTD if that limit is reached. Similarly, the simulation always evaluates a full cohort of patients, whereas in practice, where patients are more likely entered sequentially than simultaneously, a 3 + 3 design might, for example, forego the last patient in a cohort of three if the first two patients had experienced DLT. Specifics of the designs used in the simulation study are given below.

3.3.2.1 Traditional 3 + 3 Design

This design is implemented as described in Section 3.2. In the event that excessive toxicity occurs at d_1 , the MTD is taken to be d_0 . Although this is an unlikely occurrence in practice, a clinical protocol should specify any provision to decrease dose if the stopping criteria are met at the first dose level.

3.3.2.2 Traditional Best-of-5 Design

Again implemented as described in Section 3.2, with the same rules applied to stopping at d_1 .

3.3.2.3 Storer’s Two-Stage Design

Implemented as described in Section 3.2, with a second-stage sample size of 24 patients. A standard logistic model is fit to the data. If it is not the case that $0 < \hat{\beta} < \infty$, then the geometric mean of the last dose level used and the dose level that would have been assigned to the next cohort is used as the MTD. In either case, if that dose is higher than the highest dose at which patients have actually been treated, then the latter is taken as the MTD.

3.3.2.4 Non-Bayesian CRM Design

We start the design using the first stage of the two-stage design as described above. Once heterogeneity has been achieved, 24 patients are entered in cohorts of three. The first cohort is entered at the same dose level as for the second stage of the two-stage design; after that successive cohorts are entered using likelihood based updating of the dose–response curve. For this purpose we use a single-parameter logistic

model—a two parameter model with β fixed at 0.75. This value does have to be tuned to the actual dose scale but is not particularly sensitive to the precise value. That is, similar results would be obtained with β in the range 0.5–1. For reference, on the natural log scale the distance $\log(d_{MTD}) - \log(d_1) \approx 2$, and the true value of β in the simulation ranges from 2.01 to 0.37 as $\psi(d_1)$ ranges from 0.01 to 0.20. After each updating, the next cohort is treated at the dose level with estimated probability of DLT closest in absolute value to θ ; however, the next level cannot be more than one dose level higher than the current highest level at which any patients have been treated. The level that would be chosen for a hypothetical additional cohort is the MTD; however, if this dose is above the highest dose at which patients have been treated, the latter is taken as the MTD.

3.3.3 SIMULATION AND RESULTS

The simulation is performed by generating 5000 sequences of patients and applying each of the designs to each sequence for each of the dose–response curves being evaluated. The sequence of patients is really a sequence of pseudo-random numbers generated to be Uniform(0,1). Each patient's number is compared to the hypothetical true probability of DLT at the dose level at which the patient is entered for the dose–response curve being evaluated. If the number is less than that probability, then the patient is taken to have experienced DLT.

Figure 3.2 displays results of the simulation study above that relate to the estimate $\hat{d}_{MTD}$. Since the dose scale is arbitrary, the results are presented in terms of $\psi(\hat{d}_{MTD})$. Panel (a) displays the mean probability of DLT at the estimated MTD. The horizontal line at 1/3 is a point of reference for the target θ . Although none of the designs is unbiased, all except the conservative 3 + 3 design perform fairly well across the range of dose–response curves. The precision of the estimates, taken as the root MSE of the probabilities $\psi(\hat{d}_{MTD})$, is shown in panel (b). In this regard the CRM and two-stage designs perform better than the best-of-5 design over most settings of the dose–response curve. One should also note that, in absolute terms, the precision of the estimates is not high even for the best designs.

In addition to the average properties of the estimates, it is also relevant to look at the extremes. Panels (c) and (d) present the fraction of trials where $\psi(\hat{d}_{MTD}) < 0.20$ or $\psi(\hat{d}_{MTD}) > 0.50$, respectively. The see-saw pattern observed for all but the two-stage design is caused by changes in the underlying dose–response curve as the probability of DLT at particular dose levels moves over or under the limit under consideration. Since the three designs select discrete dose levels as $\hat{d}_{MTD}$, this will result in a corresponding decrease in the fraction of estimates beyond the limit. The cutoff of 0.20 is the level at which the odds of DLT are half that of θ . Although this may not be an important consideration, to the extent that the target θ defines a dose with some efficacy in addition to toxicity, the fraction of trials below this arbitrary limit may represent cases in which the dose selected for subsequent evaluation in efficacy trials is “too low”. Because of their common first-stage design that uses single patients at the initial dose levels, the two-stage and CRM designs do best in this regard. Conversely, the cutoff used in panel (d) is the level at which the odds of toxicity are twice that of θ . Although the occurrence of DLT in and of itself is not

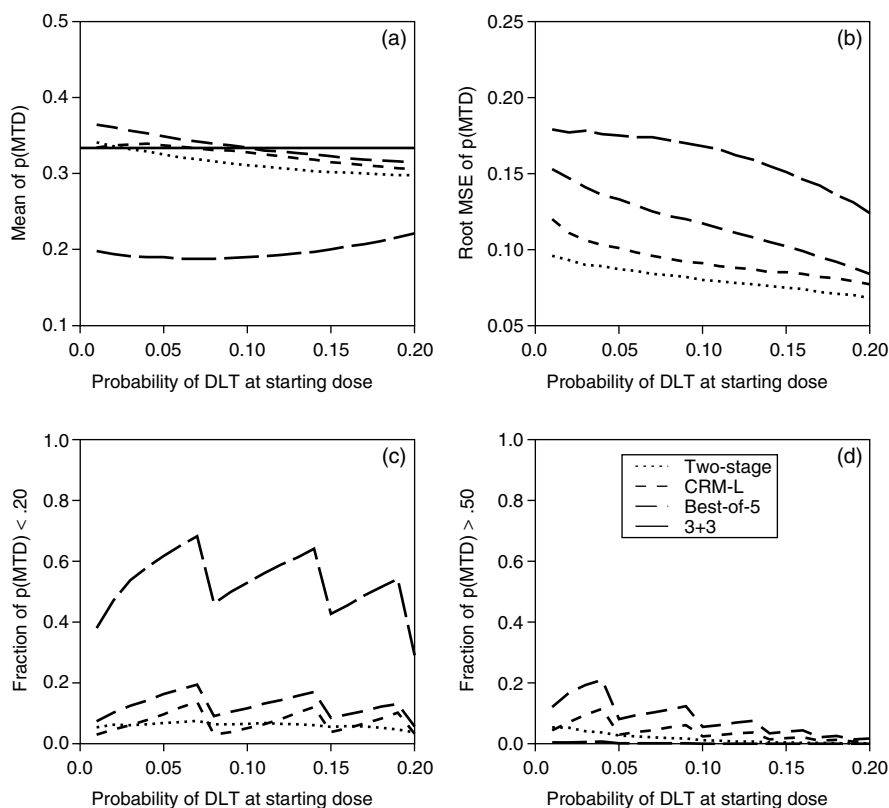


FIGURE 3.2 Results of 5000 simulated phase I trials according to four designs, plotted as a function of the probability of DLT at the starting dose level. The true MTD is fixed at four dose levels above the starting dose, with $\theta = 1/3$. Results are expressed in terms of $p(MTD) = \psi(\hat{d}_{MTD})$.

necessarily undesirable, as the probability of DLT increases there is likely a corresponding increase in the probability of very severe or even fatal toxicity. Hence, the trials where the probability of DLT is above this arbitrary level may represent cases in which the dose selected as the MTD is “too high”. In this case there are not large differences among the designs, and in particular we find that the two designs that perform the best in panel (c) do not carry an unduly large penalty. One could easily evaluate other limits if desired.

Some results related to the outcome of the trials themselves are presented in Figure 3.3. Panels (a) and (b) present the overall fraction of patients that are treated below and above, respectively, the same limits as for the estimates in Figure 3.2. The two-stage and CRM designs perform best at avoiding treating patients at the lower dose levels; the two-stage design is somewhat better than the CRM design at avoiding treating patients at higher dose levels, although of course it does not do as well as the very conservative 3 + 3 design.

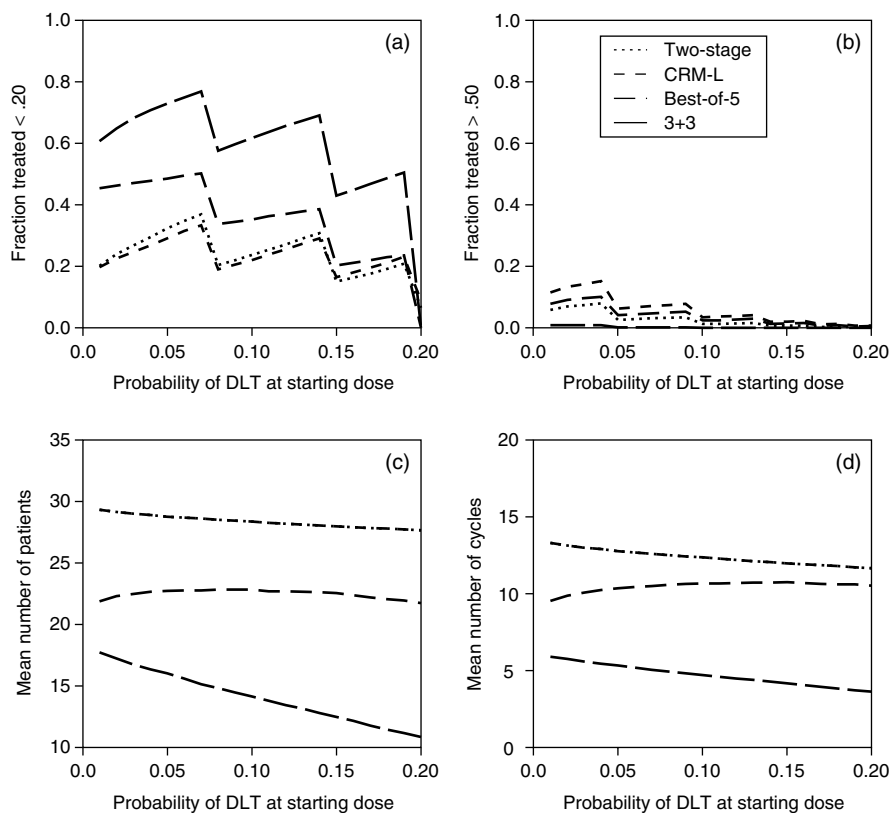


FIGURE 3.3 Results of 5000 simulated phase I trials according to four designs, plotted as a function of the probability of DLT at the starting dose level. The true MTD is fixed at four dose levels above the starting dose, with $\theta = 1/3$.

Sample size considerations are evaluated in panels (c) and (d). Panel (c) shows the mean number of patients treated. Because they share a common first stage and use the same fixed number of patients in the second stage, the two-stage and CRM designs yield identical results. The 3 + 3 design uses the smallest number of patients, but this is because it tends to stop well below the target. On average, the best-of-5 design uses 6–8 fewer patients than the two-stage or CRM design. Panel (d) displays the mean number of “cycles” of treatment that are needed to complete the trial, where a cycle is the period of time over which a patient or group of patients needs to be treated and evaluated before a decision can be made as to the dose level for the next patient or group. For example, the second stage in the two-stage or CRM designs always uses 8 cycles; each dose level in the 3 + 3 design uses either 1 or 2 cycles, etc. This is a consideration only for situations where the time needed to complete a phase I trial is not limited by the rate of patient accrual but by the time needed to treat and evaluate each group of patients. In this case the results are qualitatively similar to that of panel (e).

3.4 SUMMARY AND CONCLUSION

Based only on the results above, one would likely eliminate the 3 + 3 design from consideration. The best-of-5 design would probably also be eliminated as well, owing to the lower precision and greater likelihood that the MTD will be well below the target. On the other hand, the best-of-5 design uses fewer patients. If small patient numbers are a priority, it would be reasonable to consider an additional simulation in which the second-stage sample size for the two-stage and CRM designs is reduced to, say, 18 patients. This would put the average sample size for those designs closer to that of the best-of-5, and one could see whether they continued to maintain an advantage in the other aspects. Between the two-stage and CRM designs, there is perhaps a slight advantage to the former in terms of greater precision and a smaller chance that the estimate will be too far above the target; however, the difference is likely not important in practical terms and might vary under other dose–response conditions. The advantage of the two-stage design may seem surprising, given that the next dose level is selected only on the basis of the outcome at the current dose level and ignores the information that CRM uses from all prior patients. However, the two-stage design also incorporates

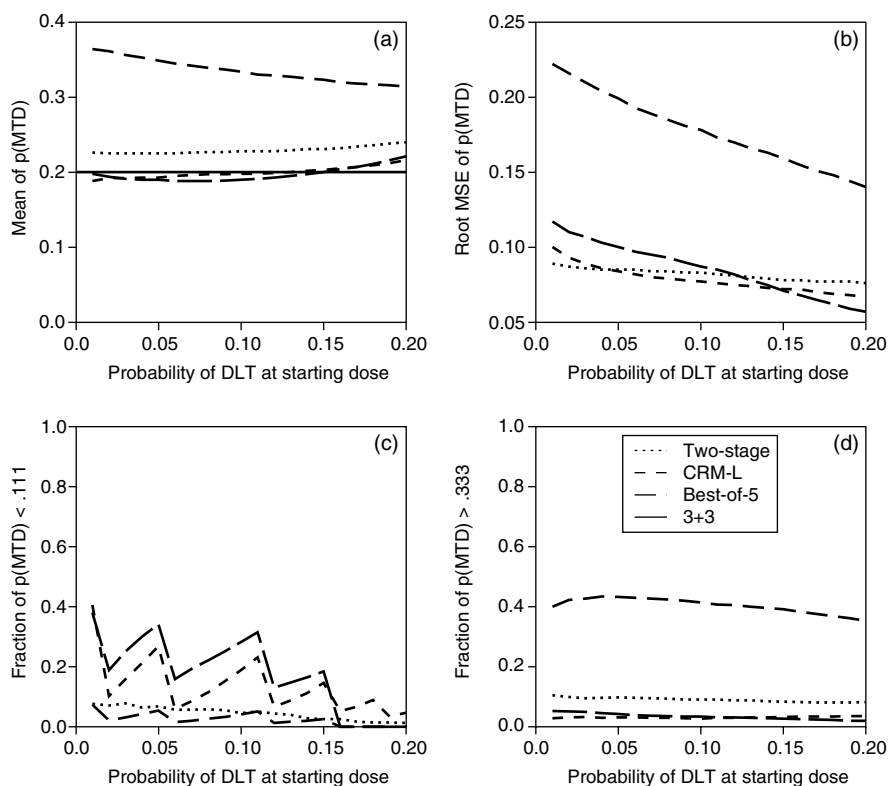


FIGURE 3.4 Results of 5000 simulated phase I trials according to four designs, plotted as a function of the probability of DLT at the starting dose level. The dose–response curves are identical to those used for Figure 3.2 but with $\theta = 0.20$. Results are expressed in terms of $p(MTD) = \psi(\hat{d}_{MTD})$.

a final estimation procedure for the MTD that utilizes all the data and uses a richer family of dose–response models. This issue is examined in Storer.¹⁶

A desirable feature of the results shown is that both the relative and absolute properties of the designs do not differ much over the range of dose–response curves. Additional simulations could be carried out that would also vary the distance between the starting dose and the true MTD or place the true MTD between dose levels instead of exactly at a dose level.

To illustrate further some of the features of phase I designs and the necessity of studying each situation on a case by case basis, we repeated the simulation study above using a target $\theta = 0.20$. Exactly the same dose–response settings are used, so that the results for the two traditional designs are identical to those shown previously. The two-stage design is modified to use five cohorts of five patients but follows essentially the same rule for selecting the next level described above with “3” replaced by “5”. Additionally, the final fitted model estimates the MTD associated with the new target; and of course the CRM design selects the next dose level based on the new target.

The results for this simulation are presented in Figures 3.4 and 3.5. In this case the best-of-5 design is clearly eliminated as too aggressive. However, and perhaps

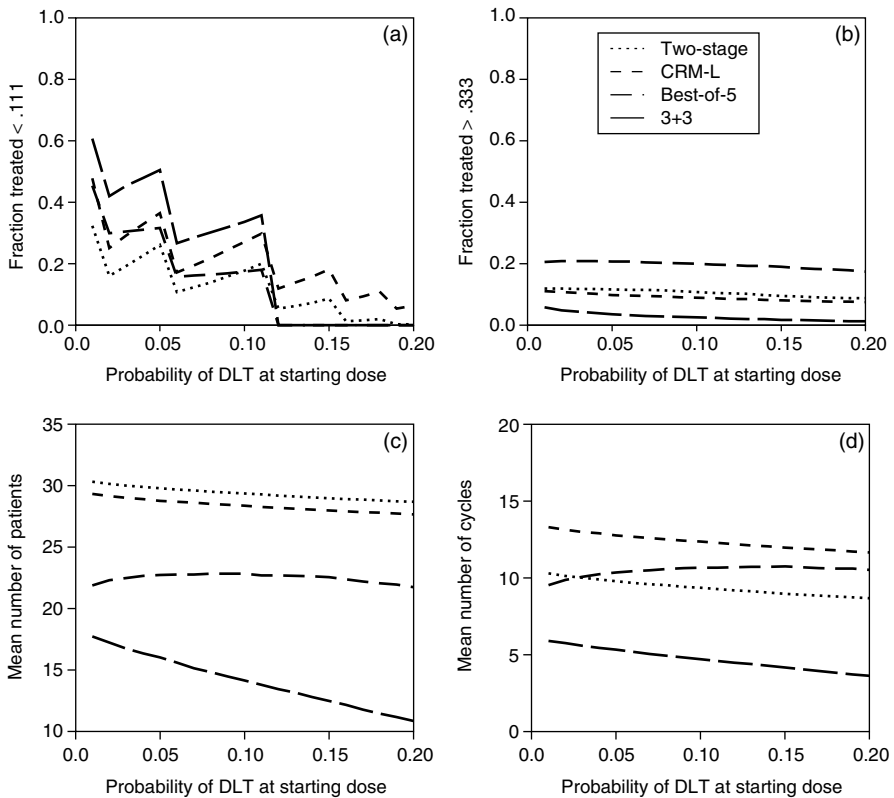


FIGURE 3.5 Results of 5000 simulated phase I trials according to four designs, plotted as a function of the probability of DLT at the starting dose level. The dose–response curves are identical to those used for Figure 3.3 but with $\psi(d_{MTD}) = 0.20$.

surprisingly, the 3 + 3 design performs nearly as well or better than the supposedly more sophisticated two-stage and CRM designs. There is a slight disadvantage in terms of precision, but given that the mean sample size with the 3 + 3 design is nearly half that of the other two, this may be a reasonable tradeoff. Of course, it could also be the case in this setting that using a smaller second-stage sample size would not adversely affect the two-stage and CRM designs.

Finally, we reiterate the point that the purpose of this simulation was to demonstrate some of the properties of phase I designs and of the process of simulation itself, not to advocate any particular design. Depending on the particulars of the trial at hand, any one of the four designs might be a reasonable choice. An important point to bear in mind is that traditional designs must be matched to the desired target quantile and will perform poorly for other quantiles. CRM designs are particularly flexible in this regard; the two-stage design can be modified to a lesser extent.

REFERENCES

1. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical studies in cancer. *Biometrics* 1990;46:33–48.
2. O'Quigley J, Chevret S. Methods for dose finding studies in cancer clinical trials: a review and results of a Monte Carlo study. *Statistics in Medicine* 1991; 10:1647–1664.
3. Shen LZ, O'Quigley J. Consistency of continual reassessment method under model misspecification. *Biometrika* 1996;83:395–405.
4. Gatsonis C, Greenhouse JB. Bayesian methods for phase I clinical trials. *Statistics in Medicine* 1992;11:1377–1389.
5. O'Quigley J, Shen LZ. Continual reassessment method: a likelihood approach. *Biometrics* 1996;52:673–684.
6. Goodman SN, Zahurak ML, Piantadosi S. Some practical improvements in the continual reassessment method for phase I studies. *Statistics in Medicine* 1995; 14:1149–1161.
7. Heyd JM, Carlin BP. Adaptive design improvements in the continual reassessment method for phase I studies. *Statistics in Medicine* 1999;18:1307–1321.
8. Piantadosi S, Liu G. Improved designs for dose escalation studies using pharmacokinetic measurements. *Statistics in Medicine* 1996;15:1142–1149.
9. Storer B. Design and analysis of phase I clinical trials. *Biometrics* 1989;45:925–937.
10. Storer B. Small-sample confidence sets for the MTD in a phase I clinical trial. *Biometrics* 1993;49:1117–1125.
11. Berlin J, Stewart JA, Storer B, Tutsch KD, Arzooonian RZ, Alberti D, Feierabend C, Simon K, Wilding G. Phase I clinical and pharmacokinetic trial of penclofenide utilizing a novel, two-stage trial design. *Journal of Clinical Oncology* 1998; 16:1142–1149.
12. Eichhorn BH, Zacks S. Sequential search of an optimal dosage. *Journal of the American Statistical Association* 1973;68:594–598.
13. Mick R, Ratain MJ. Model-guided determination of maximum tolerated dose in phase I clinical trials: evidence for increased precision. *Journal of the National Cancer Institute* 1993;85:217–223.
14. Anbar D. Stochastic approximation methods and their use in bioassay and phase I clinical trials. *Communications in Statistics: Theory and Methods* 1984; 13:2451–2467.
15. Babb J, Rogatko A, Zacks S. Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine* 1996;17:1103–1120.

16. Storer B. An evaluation of phase I designs for continuous dose response. *Statistics in Medicine* 2001;20:2399–2408.
17. Korn EL, Midthure D, Chen TT, Rubinstein LV, Christian MC, Simon R. A comparison of two phase I trial designs. *Statistics in Medicine* 1994;13:1799–1806.
18. Ahn C. An evaluation of phase I cancer clinical trial designs. *Statistics in Medicine* 1998;17:1537–1549.

4 Pharmacokinetics in Clinical Oncology: Statistical Issues

Gary L. Rosner, Peter Müller, Simon Lunagomez, and Patrick A. Thompson

CONTENTS

4.1	Introduction	77
4.2	Pharmacokinetics and Pharmacodynamics of Anticancer Agents	79
4.2.1	Example	82
4.2.2	Nonparametric Model	85
4.2.3	Nonparametric Example	87
4.2.4	Combining Data	87
4.2.5	Dose Individualization	87
4.2.6	Dose Individualization Example	88
4.2.7	Design	90
4.2.8	Design Example	90
4.3	Summary	91
4.4	Software	92
	Acknowledgments	92
	References	92

4.1 INTRODUCTION

All drugs share the feature that they are formulated to have an effect on some body system. In oncology, that effect may be to shrink the size of a tumor, reduce the growth rate of the tumor, or protect noncancer cells from potentially harmful effects of chemotherapy, to name a few examples. Pharmacokinetics (PK) is the study of what happens to drugs once they enter the body. The study of a drug's PK often will entail drawing blood samples to measure the concentration of the compound, possibly along with metabolites, over time. In some instances, it is even possible to measure the amount of the drug or metabolite in the tumor tissue itself, such as in leukemia, or in the microenvironment in which the cancer resides, such as through the use of microdialysis.¹

It is generally not enough to know the pharmacokinetics of the drug, since we are most interested in the clinical effect of the drug. Pharmacodynamics (PD) is the study of how pharmacokinetics relates to measured outcomes, such as clinical response, risk of toxicity, or even some *in vitro* measure of cellular response to a drug. A perhaps overly simplistic, though useful, characterization of pharmacokinetics and pharmacodynamics is that PK measures what happens to the drug after it enters the body and PD measures what happens to the body after the introduction of the drug.

As with other aspects of biomedical science and clinical research in the beginning of the 21st century, researchers have begun to look into genetic variation and associations with the PK and PD of drugs. Although there do not yet appear to be firm definitions, the terms *pharmacogenetics* and *pharmacogenomics* describe such studies. We will differentiate these two areas of study with the following definitions. Pharmacogenetics studies the heritable factors that contribute to PK and PD variation in a population. That is, how does genetic variation within and between populations contribute to clinical differences in how individuals react or respond to drugs? Pharmacogenomics, on the other hand, examines how mutations and other differences at the level of the genome might provide useful targets for drug development. The National Institutes of Health of the United States is supporting a large effort to aid research into pharmacogenetics and pharmacogenomics.²

The statistical issues that arise in pharmacogenetic studies are similar to the issues that arise in most genetic epidemiology studies. For example, if one has identified a gene associated with the metabolism of an anticancer drug and one wishes to see if genetic variation at this allele is associated with different risks of toxicity, one may wish to carry out a study prospectively or retrospectively. The prospective study design might call for blood samples at entry to the study, with genotyping. Then one would look for associations between genotype and whether or not a patient experienced toxicity. For a retrospective study, one might identify patients who have experienced toxicity and find an appropriate control group, perhaps with matching. Then one would only have to genotype this subset of the study's patients for analysis of the association between toxicity and genotype. Obviously, if some genetic variants put the patient at risk of fatal toxicity, the retrospective design may well miss these patients without banked blood samples.

Some of the other statistical issues that arise in pharmacogenetic studies include single nucleotide polymorphism (SNP) studies and determination of haplotypes. With the many variants at specific loci within a gene, there exists the potential for carrying out many hypothesis tests and declaring false positive associations. Furthermore, many SNPs are in linkage disequilibrium, meaning they tend to move together and be highly correlated. One can gain some extra power by grouping SNPs together if they are highly correlated with each other. Statistical techniques to group SNPs into such groups, called haplotype blocks, are an endeavor carried out in pharmacogenetic studies as well as other genetic-based studies. More information is contained in the chapter on haplotypes in this volume. Also, information about pharmacogenomics in general and oncology-specific issues is contained in a recent edited volume.³

In this chapter, we will explore some of the main statistical issues in the study of pharmacokinetics and pharmacodynamics in clinical oncology. Most of these same issues arise in clinical research relating to other diseases, but our focus is cancer.

4.2 PHARMACOKINETICS AND PHARMACODYNAMICS OF ANTICANCER AGENTS

The most common basis for determining a drug's pharmacokinetics is measured drug concentration in the blood. PK studies call for drawing blood samples during and/or after an infusion and assaying the concentration of the compound in the blood. The relationship between the drug's concentrations and time is characterized by a system of differential equations that defines a pharmacokinetic model.⁴ Quite often, the differential equations incorporate so-called compartments to allow for changes in the loglinear decay of the drug after the end of the infusion. These differential equations describe the instantaneous change in concentration of the drug or its metabolites within each compartment, with direct or indirect communication between compartments.

We often use compartmental models to characterize the relationship between plasma concentration and time, primarily because they seem to fit and have a physiologic interpretation. The compartments in these models are based on the notion that the drug circulates through the body in the blood stream and may visit other parts of the body before it is eliminated. The plasma or blood compartment may be considered the central compartment for a drug that is infused intravenously, but the drug will likely pass through the liver, kidneys, and, hopefully, the tumor. Eventually, most of the drug returns to the compartment from which it is eliminated, which may be the central compartment. The transit between the plasma and these other organs, or compartments, forms part of the system of differential equations characterizing the change in concentration over time. Figure 4.1 illustrates a simple two-compartment model.

Depending on the form of the differential equation describing the instantaneous rate of change of a compartment's concentration, the kinetics may be linear or nonlinear. A drug is said to have linear kinetics if for two different doses the concentrations that result from each dose are proportional to the doses. If, however, the concentrations arising from different doses are not proportional to doses, then the drug exhibits nonlinear kinetics. Nonlinear kinetics can occur if a process gets saturated so that no matter what the concentration the rate of change remains constant.

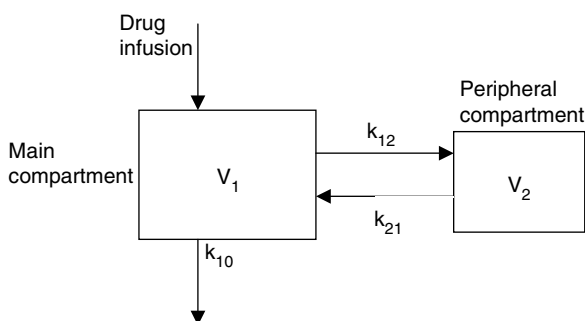


FIGURE 4.1 Simple two-compartment model characterizing the disposition of a drug given by continuous infusion.

At the risk of oversimplification, the drug exhibits linear kinetics if one can take the concentration–time curve associated with one dose of the drug, rescale, and perhaps shift it and end up with the concentration–time curve associated with another dose. Modeling nonlinear kinetics is more difficult computationally because there is no analytic solution to the differential equations. Instead, one has to find the solution numerically.

Generally, one has to use a nonlinear regression package, such as the `nls` function in R, to estimate an individual's model parameters. (There are software packages specifically for fitting PK models; please see the end of this chapter for a list of several noncommercial packages for fitting pharmacokinetic models.) In the compartmental model, the unknown parameters to estimate include the volumes of the compartments and the rate constants for the movement of drug between compartments and out of the body. The volume of distribution, labeled V_1 in Figure 4.1, is the hypothetical volume of the main compartment from which we measure the concentrations over time. Being a parameter in the PK model, the volume of distribution is hypothetical in the sense that it is generally not interpretable as the volume of plasma, say, in the body. Patient-specific estimates of volumes of distribution often exhibit greater variability and sometimes much larger values than one would expect for the volume of blood in human bodies. In essence, the compartment-specific volumes of distribution and rate parameters are model parameters that have a theoretical interpretation loosely based on physiologic considerations.

Some of the basic parameters that describe the pharmacokinetics of a drug include clearance, area-under-the-concentration–time curve (AUC), steady-state concentration, the elimination half-life, volume of distribution, the elimination rate parameter, and bioavailability for drugs administered via a route other than directly into the blood stream. These characteristics of the pharmacokinetics of a drug are either direct parameters in the concentration–time function or are functions of the model parameters. Clearance, measured in volume per unit time (e.g., liters per hour), is most often used to characterize a drug or to compare a drug's pharmacokinetics in different populations. By clearance, one generally means *total body clearance*, that is, the sum of the individual clearances of the drug from all organs. Another important measure is the AUC, which is a measure of systemic exposure to the drug. The AUC is usually given in units of concentration times time (e.g., $\mu\text{g} \times \text{hour}/\text{ml}$). If a drug is given by intravenous infusion, clearance (CL) and AUC are inversely proportional, with the proportionality being the infused dose. That is, $\text{CL} = d_{\text{iv}}/\text{AUC}$, where d_{iv} is the administered dose.

Many anticancer drugs are given by constant continuous intravenous infusion over an extended time. During the infusion, plasma concentrations of the drug increase but eventually reach a constant value if the infusion lasts long enough. The steady-state concentration (C_{ss}) is, as the name implies, the plateau concentration reached in the plasma during continuous infusion at a constant rate. Steady-state concentration and clearance are inversely proportional to each other. The proportionality constant is the infusion rate, i.e., $\text{CL} = \text{InfRate}/C_{\text{ss}}$.

The elimination half-life is the time it takes for half of the administered dose to be eliminated from the body. Usually, the plasma concentration is within 10% of the C_{ss} after an infusion lasting four half-lives of the drug.

These PK-derived parameters are functions of the compartmental model's parameters. Thus, once one has fit the model, one can estimate these derived parameters. There are two exceptions. As indicated above, one can estimate total-body clearance if one has an estimate of the steady-state concentration during a continuous infusion at a constant infusion rate. The other exception is a so-called noncompartmental estimate of the AUC. If one has enough concentration–time pairs for a patient, one can use the trapezoidal rule to compute the area under the concentration–time curve. If the final concentration is not zero, a reasonable approximation might be a simple exponential decay with slope based on the logarithm of the last two concentrations. This estimate for the AUC is attractive because it does not depend on any model assumptions. Usually, however, one sees the estimated AUC without any attempt to attach a measure of uncertainty to the estimate.

Many anticancer drugs are available in oral form, which is more convenient in case of chronic dosing or for patients who may find it difficult to get to a clinic. When drugs are taken orally, they pass through and are disseminated from the gut. Ultimately, the drug reaches the circulation system, from which it will get to the rest of the body. Bioavailability relates the amount of drug in the blood stream after oral administration to the amount of the drug in the blood stream after intravenous administration. Bioavailability will typically be between 0 and 1. Bioavailability, which has to be estimated, contributes to PK and PD variability.

The disposition of a drug in the plasma or tissue may be affected by many factors including genetics, environment, diet, age, and other drugs being taken or foods being digested at the same time the drug is in the body. As part of studying a drug's pharmacokinetics, one will often look for associations between patient characteristics and patient-specific PK parameters. A common approach is to fit separate models to each patient's concentration–time data and then carry out statistical inference on the patient-specific PK model parameter estimates. This two-stage approach ignores uncertainty in the parameter estimates and may thereby lead to false declarations of significant differences. For example, one might regress each patient-specific AUC or clearance on age or smoking status or genotype to look for patterns or potentially significant differences. Analyses of patient-specific model estimates and covariates are exploratory because they typically ignore the uncertainty in the PK-derived parameters. Inference that accounts for all sources of uncertainty is preferable.

Pharmacokinetics modelers have long realized that there is heterogeneity between individuals in terms of subject-specific model parameters. This realization led to the use of mixed-effects models and hierarchical modeling, allowing for between and within variation. Sheiner and colleagues^{5–7} were among the first to recognize the usefulness of these models for predicting the time course of a drug's concentration for individuals. Other researchers followed up on these ideas, which led to a general approach to studying variation of the pharmacokinetics of a drug in a population called population modeling.⁸ Population modeling is, essentially, a hierarchical model, as in Equation 4.1:

$$\begin{aligned} y_{ij} &= f(t_{ij}, \theta_i) + e_{ij}, e_{ij} \sim F \text{ with } E[e_{ij}] = 0 \\ \theta_i | \theta_0 &\sim G(\theta) \text{ with } E[\theta_i] = \theta_0 \\ \theta_0 &\sim H(\bullet) \end{aligned} \tag{4.1}$$

In this equation, y_{ij} is the concentration of the drug at time t_{ij} for the i^{th} patient. The patient's own model parameters are denoted by θ_i , which are assumed distributed in the population. The population distribution of these individual parameters θ_i is characterized by G and indexed by parameter θ_0 , corresponding to the mean of the θ_i in Equation 4.1. We express our uncertainty about the mean parameter value in the population through the hyperprior distribution denoted H . The difference between the measured concentrations for the i^{th} patient at time t_{ij} and the modeled concentrations is a random variable e_{ij} having zero mean and distribution F .

Typically, the distribution of the residuals is considered normal or lognormal. The variance may be a function of the mean concentration level, such as when one wishes to fit a constant coefficient of variation model. The distribution of subject-specific model parameters is also often treated as normal or lognormal. A frequentist analysis would typically stop at the level of the population distribution. A Bayesian analysis, in contrast, will also specify a distribution for the parameters in the population distribution via a hyperprior distribution. The inclusion of a hyperprior and hyperparameters allows for between-individual variation while providing information about the distribution of subject-specific parameters. A Bayesian hierarchical model also makes prediction for a new patient straightforward. The book by Davidian and Giltinan describes well many of the frequentist and Bayesian methods for analyzing repeated measures having a nonlinear relationship with time.⁹

4.2.1 EXAMPLE

As an example, we consider the analysis of the anticancer agent methotrexate in infants (children less than one year old) with leukemia. The clinical prognosis for children younger than one year old who are diagnosed with acute lymphoblastic leukemia (ALL) is worse than for older children. It may be that the disease is different or, perhaps, the infant's organs, particularly the kidneys, are not yet fully developed causing the pharmacokinetics of anticancer drugs to be different in the infants.

Children with ALL often receive methotrexate as part of their chemotherapeutic treatment regimen. Methotrexate is cleared from the body by the kidneys, and variability in the drug's pharmacokinetics may be associated with key measures of renal function, such as glomerular filtration rate (GFR), tubular secretion, and renal blood flow. Each of these characteristics of renal function changes as the kidneys mature during the first few months after birth. Additionally, the liver may continue to develop after birth, leading to altered drug metabolism during the first few months of life. Thus, as infants develop, there may be changes in the drug's absorption, distribution, metabolism, and elimination.

Little is known about the pharmacokinetics of methotrexate in very young children. Therefore, pediatric oncologists were interested to learn what they could by collecting data as part of a larger study carried out by the Children's Oncology Group (COG). In this study, infants with ALL received methotrexate as a high-dose infusion (4 g/m²) over 24 h on weeks 4, 5, 11, and 12 from the start of the treatment regimen. The infusion was divided into an initial 200 mg/m² loading dose given over 20 min followed by 3.8 g/m² infused over 23 h and 40 min, the remainder of the 24 h infusion duration. The patients received 24 h infusions of methotrexate a week apart and again seven and eight weeks after the first infusion of the drug.

The PK study combined data from two sources. As part of routine monitoring of the infants on the study, some blood samples allowed the measurement of methotrexate concentration levels. The patient charts held these data, allowing for retrospective collection. The remaining patients in our sample underwent more extensive sampling. Aside from the dose of methotrexate actually received and the concentration–time data, the dataset also included patient characteristics such as age, height, weight, body-surface area, and serum creatinine. The protocol called for measuring methotrexate levels in each course of therapy at the end of the 24 h infusion and every 24 h following the end of the infusion until the concentration was less than 0.18 μM . The eighteen patients who underwent more extensive sampling had their additional blood draws during their first methotrexate infusion. Sample collection for this subset of the patients was at 1, 6, 12, and 23 hours after the start of the first methotrexate infusion.

Altogether, the dataset included seventy patients with enough information to model at least one course of methotrexate. We could analyze data during the first course of therapy for 62 patients. The dataset contained a total of 686 methotrexate concentration–time pairs measured during 199 doses of methotrexate given to the seventy infants.

The primary measure of kidney function in the dataset was an estimate of the glomerular filtration rate (GFR). The estimate is a function of the infant's length and the creatinine measured in the infant's plasma.^{10,11}

Figure 4.2 shows the concentration–time data for two of the more extensively sampled patients during their first course of methotrexate. One sees in the figures how the concentrations approach a plateau corresponding the steady-state concentration over the course of the 24 h infusion. The figures also show that these two patients have quite different steady-state concentrations (around 52 $\mu\text{moles}/(\text{ml} \times \text{m}^2)$ vs. around 35 $\mu\text{moles}/(\text{ml} \times \text{m}^2)$).

As indicated above, one can estimate the total-body clearance of a drug from the steady-state concentration and the infusion rate. Even though the infusion rate was not constant over the 24 h infusion, it was constant after the first twenty minutes. Therefore, it seemed reasonable to estimate the clearance at steady state from these data.

If one estimates clearance from the end-of-infusion or 23 h and 24 h concentrations and uses these estimates in a regression analyses with patient covariates, then one is ignoring the uncertainty associated with these estimates. Ignoring this uncertainty may result in a false-positive finding of an association between clearance and patient characteristics. Instead, we chose to fit a population or hierarchical model. We used the program PKBugs,¹² which runs in WinBUGS.¹³ We fit a model for the log concentrations having a normal residual, with the subject-specific parameters also being lognormal. The full hierarchical model is

$$\log(y_{i,j}) \sim N(\log(C[t_{i,j}, \theta_i, d_i])\tau)$$

$$\theta_i \sim MVN(X_i\psi, \Omega)$$

$$\psi \sim MVN(\psi_0, C)$$

$$\tau^{-1} \sim \text{Gamma}(0.001, 0.001), \Omega^{-1} \sim \text{Wishart}(R, \rho)$$

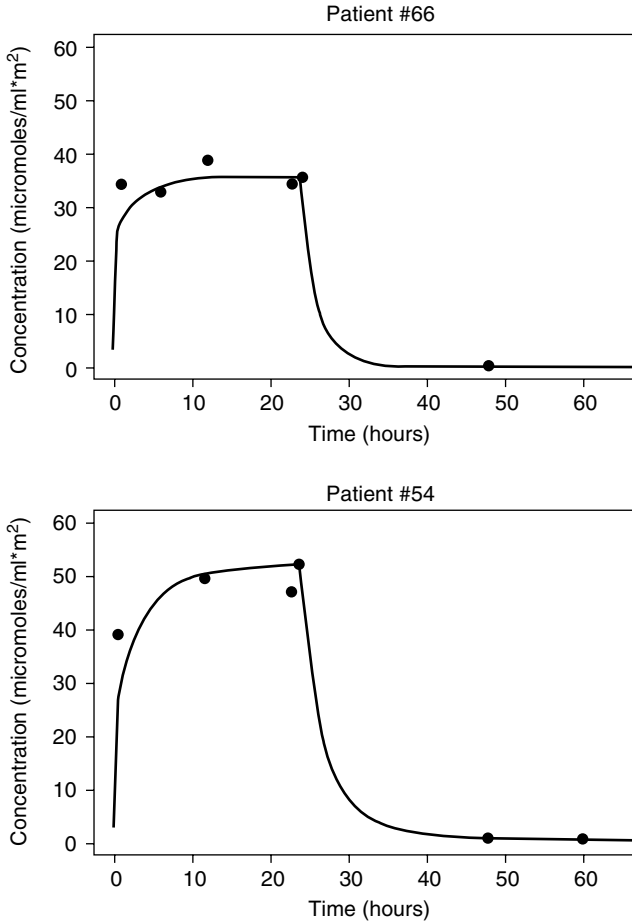


FIGURE 4.2 Observed and fitted methotrexate concentrations for two infants in the study. The lines are based on point-wise posterior means.

Here, $y_{i,j}$ is the concentration measured on patient i at time $t_{i,j}$. In the model, $N(\mu, \tau)$ denotes a normal distribution with mean μ and variance τ , while a multivariate normal distribution with mean vector $\mathbf{m}$ and variance–covariance matrix Ω is denoted $MVN(\mathbf{m}, \Omega)$. The compartmental model is denoted $C(t_{i,j}, \theta_i, d_i)$, a function of time, model parameters, and dose d_i . The function relating concentration to time, parameters, and dose for a two-compartment model (see Figure 4.1) can be written as a sum of two exponential functions.^{4,12}

$$C(t_{i,j}, \theta_i, d_i) = A_1 e^{-\lambda_1 t} (1 - e^{\lambda_1 t^*}) + A_2 e^{-\lambda_2 t} (1 - e^{\lambda_2 t^*})$$

T = infusion duration

$$t^* = \min(t, T)$$

$$A_i = \frac{k_0(k_{21} - \lambda_i)}{V_1\lambda_i(\lambda_{3-i} - \lambda_i)}, \quad i = 1, 2$$

$$k_0 = \text{infusion rate} = \frac{\text{dose}}{\text{infusion duration}}$$

The two parameters λ_1 and λ_2 are functions of the rate parameters shown in Figure 4.1. They satisfy the following two equations in terms of the rate constants.

$$\lambda_1 + \lambda_2 = k_{10} + k_{12} + k_{21} \quad \text{and}$$

$$\lambda_1\lambda_2 = k_{10}k_{21}, \quad \text{with } \lambda_1 > \lambda_2 \text{ by definition}$$

The patient-specific model parameters are θ_i . This vector includes the four parameters in the pharmacokinetic model, namely, $\log(\text{Cl})$, $\log(Q)$, $\log(V_1)$, and $\log(V_2)$, recalling that each patient has his or her own set of parameters. Here, V_1 and V_2 are the respective volumes of the central and peripheral compartments shown in Figure 4.1. The clearance (Cl) equals V_1 times k_{10} . The parameter Q is called the distributional clearance between the two compartments, which is assumed to be the same. That is, Q satisfies $Q = k_{12}V_1 = k_{21}V_2$.

Returning to the hierarchical model, the hyperparameters had the following values. The variance matrix for the patient-specific parameters (Ω) had an inverse-Wishart prior with $\rho = 4$ degrees of freedom and scale matrix R corresponding to 50% coefficient of variation for each parameter. The parameters, in turn, may depend on patient characteristics X_i (e.g., the age or GFR for child i) through a linear regression with coefficients ψ . These regression coefficients had a multivariate normal prior with variance matrix C equal to 10^4 times the identity matrix. The hypermean for the regression coefficients was a vector with nonzero values for the PK parameter-specific intercepts and zero for the covariate effects. The nonzero hypermeans came from an analysis of related data.

Figure 4.2 shows the PKBugs fits to the data as the solid line. These are the piecewise posterior mean concentrations for the patients over time. Having fit the model, we examined potential relationships with covariates by plotting the posterior means of the parameters or functions of the parameters against the potential covariates.¹⁴ We did not find a strong relationship between age (0–12 months) and clearance given the large amount of noise in these data. Nevertheless, since the primary question motivating this study concerned an association between age and clearance among the infants, we modeled the logarithm of clearance as a function of age in the hierarchical model. The posterior mean of the coefficient was 0.034 (posterior standard deviation = 0.038). We did not find any strong associations between the covariates available to us and PK in this dataset. Further modeling based on exploratory plots mentioned in the previous paragraph suggested an association between the model-derived parameters GFR and clearance.

4.2.2 NONPARAMETRIC MODEL

In the previous example, we assumed a parametric distribution for the subject-specific parameters in the model. Assuming a normal or lognormal distribution

may be restrictive in some cases. Some investigators have sought to incorporate nonparametric estimation of the population distribution.^{15–18} Bayesian nonparametric modeling removes the assumption that subject-specific PK parameters vary in a population according to a normal distribution. In particular, knowledge that potential effect modifiers exist, such as genetics, different diets, etc., means that there might be multimodal distributions of the PK parameters in a study population. The earliest attempts to be nonparametric in the population distribution built up mass at discrete locations according to the data. In this approach,¹⁵ the distribution of θ_i in Equation 4.1 is left completely unspecified (i.e., $\theta_i \sim F$). The likelihood function becomes a function of the unknown distribution function F , and the problem becomes one of maximum likelihood estimation of a mixing distribution.^{19,20} One can show that the maximum likelihood estimate is a discrete distribution on at most n points of support, where n is the number of patients in the sample.^{15,19}

If one wants to include continuous covariates in the population model, one has to finesse the problem a bit to make it work with the nonparametric approaches discussed in the previous paragraph.²¹ Furthermore, one would typically consider modeling the model parameters with a continuous distribution. Davidian and Gallant used smooth nonparametric maximum likelihood to allow for a family of continuous distributions that can incorporate continuous covariates explicitly. The smooth nonparametric maximum likelihood solution to the problem estimates the underlying distribution of a k -variate random vector from a class of densities that are at least $k/2$ times differentiable. A density in the specified class may be represented for all practical purposes as a series expansion of a polynomial times a normal density function.^{22,23}

None of these nonparametric or semiparametric methods are Bayesian, however. Instead, they are mixed-effects models with unspecified distributions for the random effects. One can, however, carry out full Bayesian inference in a hierarchical model with nonlinear regression functions and still be nonparametric. Walker and Wakefield²⁴ used a Dirichlet Process²⁵ prior for the distribution of the subject-specific parameters in a hierarchical model with a nonlinear regression, such as a compartmental PK model. The posterior, however, is necessarily discrete, a result of the Dirichlet process prior.

We have found that a Dirichlet process mixture allows for nonparametric and semiparametric modeling for these problems.^{26–28} Simply put, the prior model for the subject-specific parameters is a mixture of normals with the weights and locations coming from a Dirichlet process hyperprior. That is, the distribution of subject-specific parameters in the population is a mixture of continuous densities, such as normal densities, with mixing (i.e., random weights) over the random locations. The Dirichlet process is the prior mixing measure. The discrete nature of the posterior from a Dirichlet process prior is not a problem because this prior is on the locations of a continuous density. Thus, the posterior distribution is continuous with this model. Covariates are incorporated via the conditional distribution of the parameters given the covariates. The result is a semiparametric regression yielding a smooth curve as a weighted mixture of linear regressions over the range of the covariates.

4.2.3 NONPARAMETRIC EXAMPLE

We first applied a Dirichlet process mixture model to a dataset consisting of white blood cell counts measured while patients are receiving increasing doses of a chemotherapeutic drug in a phase I study.^{28–31} For cytotoxic cancer chemotherapy, myelosuppression (lowered blood counts) is a common side effect, getting more severe as doses increase. Thus, there is great interest in monitoring patients' blood counts as a pharmacodynamic end point, especially when escalating doses in a phase I study.

The analysis of blood count data for cancer patients receiving myelosuppressive doses of chemotherapy is an example of a pharmacodynamic analysis of a nonlinear model for repeated measurement data. We implemented the Dirichlet process mixture model to analyze these data with the dose of the drug serving as the pharmacologic covariate. The model can also be fit to other repeated measurements, including drug concentrations measured over time.

Examples of non-Bayesian analyses of nonlinear models for repeated measurements data in the pharmacodynamic literature also analyzed the time-course of blood count data over time. Karlsson and colleagues³² used a spline-based model to analyze similar data, although with splines they had to constrain each patient's profile to return to baseline at the end of the course of chemotherapy. There are also models that incorporate PK and PD simultaneously by hypothesizing an *effect compartment*.^{33–35} These analyses via a so-called indirect-response model have not yet been studied much in the statistical literature. Minami and colleagues^{36,37} used this approach to model blood count data collected on cancer patients receiving different chemotherapeutic drugs.

4.2.4 COMBINING DATA

Another statistical issue in studies of population pharmacokinetics of drugs concerns combining data from multiple sources. Combining analyses across different population pharmacokinetic or pharmacodynamic studies would seem to be a good way to learn more about the distribution of PK parameters in the general population or to allow for more precise inference on the effects of patient characteristics on PK. With mixture priors, it is not at all obvious how to combine data in a sensible way that leads to borrowing strength across the studies but still allows for each study to maintain its own idiosyncrasies as characterized through the flexible mixture model. We have developed a method for use with finite mixtures allowing for a common measure and study-specific mixtures.³⁸ We have also developed an ANOVA-like decomposition of the random locations in a Dirichlet process via the dependent Dirichlet process.³⁹ This appears to be a useful modeling strategy for such meta-analyses because they maintain flexibility in the inference and allow for different degrees of exchangeability.

4.2.5 DOSE INDIVIDUALIZATION

With the ability to model a drug's pharmacokinetics and simultaneously account for between-individual variation came the realization of the potential to predict a

person's individual pharmacokinetics if given the drug. This realization then led to the possibility of PK-guided dosing, meaning that physicians and clinical pharmacologists can tailor the dose of a drug for an individual to that individual's own ability to handle the drug. Several researchers called for PK-guided dosing in medical oncology, recognizing that the often narrow therapeutic window (i.e., the narrow range of doses that are neither too toxic nor too low to allow for clinical efficacy) might be made wider if patients received doses that would be expected to lead to systemic exposure in some target range. Other areas of clinical medicine benefited from the use of PK-guided dosing, including the use of antibiotics.^{40–42} Several attempts to use PK to guide dosing in cancer have also appeared but have not yet been put into practice.^{43,44}

4.2.6 DOSE INDIVIDUALIZATION EXAMPLE

An example of a fully Bayesian design for pharmacologically-guided dosing in cancer is given by a study from the University of Texas M.D. Anderson Cancer Center. Our clinical collaborator treats patients with leukemia using bone marrow transplantation. With this form of anticancer therapy, the patient receives ultra-high doses of chemotherapy. The drugs are highly toxic, and at these doses the patient's blood cells are virtually wiped out. Without circulating white blood cells, people are subject to potentially fatal infection from pathogens that would otherwise not cause much of a reaction. In order to help the patient's body recover its ability to produce white blood cells, the transplant oncologist infuses either bone marrow or peripheral blood stem cells. An allogeneic transplant is one in which the patient receives cells from a donor who matches the patient in some way. Autologous transplants reinfuse cells removed from the patient prior to the ultra-high-dose chemotherapy.

In transplant therapy, the medical oncologist seeks to treat the patient with doses that are high enough to kill any and all cancer cells but not so high that the drugs kill the patient. Most transplant regimens give the drugs at fixed doses that are a function of body size as measured by the body-surface area.⁴⁵ If there are sources of pharmacokinetic and pharmacodynamic variation beyond body size, some patients may well receive a dose that is either too high or too low.

One might instead define a target range of exposure to the drug, such as via the area-under-the-concentration–time curve, or AUC. In fact, our clinical collaborator told us that he had a target range for the AUC of the drug busulfan when given intravenously. He wanted to design a study in which patients first received a small, non-therapeutic dose of the drug, with blood draws to allow pharmacokinetic analysis. With the pharmacokinetic model fit to the concentration–time data for the test dose, he thought that one could predict the dose that would achieve the desired AUC. We agreed and set about to design this study.

We chose to determine a fully Bayesian design for the study. There was clearly a loss function, namely achieving an AUC that is either too low or too high, with greater loss as the AUC was farther away from the ends of the target interval. Furthermore, the study called for prediction, for which the Bayesian approach excels. Among other advantages, Bayesian modeling offers the ability to incorporate sources of uncertainty in the algorithm. Finally, we already had data from earlier

studies without the test dose. Thus, we could incorporate historical information as prior information to improve precision.

We did not assume that the patients in the two historical studies were fully exchangeable, choosing instead to keep the studies separate in a hierarchical model. We did, however, want to allow for a Bayesian nonparametric prior distribution in our inference. Therefore, we chose to use the so-called dependent Dirichlet process prior in our model.³⁹ This model allows for some borrowing of strength across the studies while still retaining study-specific differences within a larger nonparametric model.

Our utility function consisted of two straight lines on either end of the target AUC range (Figure 4.3(a)). The slopes of the two lines differed in that the loss would rise more steeply for exceeding the range than for falling short of it. The reason for the lack of symmetry is that too high a dose might lead to death, whereas too low a dose might still provide some therapeutic benefit to the patient.

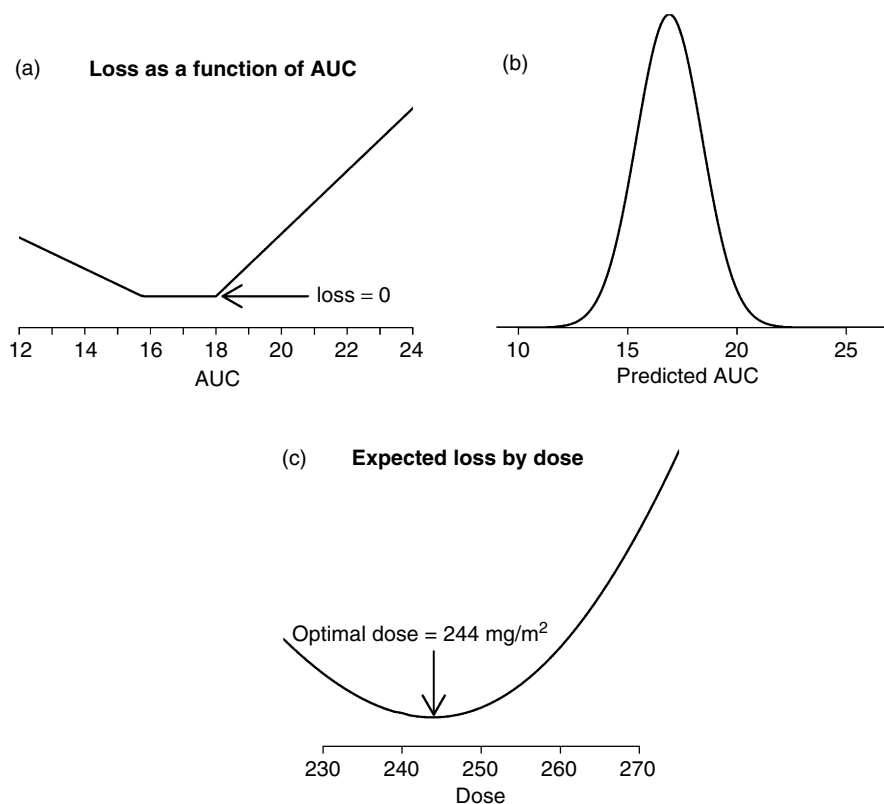


FIGURE 4.3 (a) The loss function as it relates to the AUC. (b) An example predictive distribution for the AUC if the next patient receives a particular dose. (c) The expected utility as a function of dose allowing one to determine the optimal dose while incorporating the many sources of uncertainty.

Based on the results of the PK analysis of the test dose along with the information contained in the previous studies, we computed the posterior predictive distribution of the AUC for a host of possible doses. Figure 4.3(b) shows an example predictive distribution for a hypothetical new patient receiving some fixed dose. The utility function $u(d, y, \theta)$ in this study is minus the loss associated with a given AUC and is a function of the dose of the drug (d), the concentrations over time (y), and the patient's PK parameters (θ). Integrating the utility function with respect to the posterior predictive distribution of the AUC for a fixed dose gives the expected loss for that dose. That is, the optimal dose d^* maximizes $\iint u(d, y, \theta) p_d(y|\theta) p(\theta|\text{Data}) dy d\theta$, where $p_d(y|\theta)$ is the sampling distribution of the future concentrations as a function of dose (d) and the PK parameters. Of course, the calculations would refer to a specific patient, but we have suppressed the subscripts for ease of presentation. Over a range of doses and associated expected utilities, one can pick the dose with the highest expected utility (or, in our case, the lowest expected loss), as illustrated in Figure 4.3(c). The study we designed will use the patient's test data to tailor the dose of the high-dose chemotherapy to the specific patient. This study is ongoing.

4.2.7 DESIGN

Another important aspect of designing population studies of pharmacokinetics and pharmacodynamics concerns the timing of measurements. Quite often, pharmacologists choose sampling times based on D-optimality criteria or similar criteria relating to minimizing some function of the variance–covariance matrix. Bayesian optimal design has generally been Bayesian versions of D-optimality and the like.^{46–48} An early example of using information theory to design a PK study is given by D'Argenio.⁴⁹

4.2.8 DESIGN EXAMPLE

We have proposed a fully Bayesian design with a loss function that incorporates inconvenience to the patient.⁵⁰ The Cancer and Leukemia Group B (CALGB) was preparing to carry out a large study of 3 h infusions of paclitaxel to treat women with metastatic breast cancer. The proposed sample size was large, and the study provided an excellent opportunity to study the pharmacology of this important anticancer drug among a large group of women. Investigators had reported that the PK parameter with the greatest association with myelosuppression, the primary toxicity, was the time during which the concentration of the drug is above some threshold level.⁵¹ Therefore, the CALGB was interested in having a good estimate of the time above the threshold concentration and wanted the sampling times to provide for a good estimate.

Estimating the time above some threshold concentration requires drawing samples around the time when one might expect the concentration to drop below the threshold. This requirement means that women would potentially have to stay in the clinic longer than clinically necessary or even return the next day in order to ensure a sample with the most information about when the concentration drops below a

threshold. A collaborating clinical pharmacologist suggested, in fact, that the study require a blood sample 24 h after the start of the infusion. The treating clinicians did not want to require that the women return to the clinic the day after the infusion, arguing that such a requirement would be too disruptive for the women participating in the study. We decided to approach the design question from a decision-theoretic standpoint by including an increasing cost for times after seven hours from the start of the infusion, which is the length of time the women would be asked to remain in the clinic for routine surveillance. That is, we wanted to find the optimal times in order to maximize the precision of estimated PK parameters (in particular, AUC and the time above a threshold concentration), accounting for the potential cost of requiring women to wait around or even return to the clinic the next day. The utility also included a gain as the posterior precision (inverse of the variance) of the PK characteristic of interest ($S(\theta)$), such as AUC or time above a threshold concentration) increased. Equation (4.2) shows the utility as a function of the times (t) and the data (y). The cost associated with a set of possible sampling times is the sum of each time's cost, and each time's cost is zero or the squared difference between that sampling time and 7 h. The parameter c calibrates a change in cost because of a sampling time with improved precision.

$$u(t, y) = \text{var}(S(\theta)|t, y)^{-1} - c \sum_{i=1}^k (t_i - 7)^2 \quad (4.2)$$

Of course, it is not straightforward to decide how to weigh estimation precision with the cost of waiting to draw samples because these are not on a common metric. Our approach called for working with the clinicians to calibrate the relative weights when it came time to implement the study design.

We used historical PK data for a prior distribution in our design. Because of the nonlinear nature of the mean function, we had to use Markov chain Monte Carlo (MCMC) methods for inference.⁵² We also used MCMC methods to determine the utility surface, treating the problem as one in which we generate random samples from some probability distribution known up to a constant of proportionality.^{53,54} In the end, we found that the utility surface was relatively flat, and we had to work with a numerically exaggerated transformation of the surface to find the times associated with the peak utility. Of course, it was highly informative for the clinical pharmacologists to know that the utility surface was relatively flat over some range of potential sampling times because this meant that there would not be much loss in using sampling times that might be *suboptimal* yet be appealing for other reasons not captured by the utility function.

4.3 SUMMARY

In this chapter, we have highlighted some of the statistical issues that arise in studies of the pharmacokinetics and pharmacodynamics of a drug. We used case studies when available to illustrate methods. The field is a fascinating one and has the potential to individualize therapy for patients, thereby maximizing the chance of clinical benefit while minimizing the risk of serious toxicity.

4.4 SOFTWARE

Commercial software packages for fitting pharmacokinetic models are available. Some programs are available free to academic institutions. Here are three software packages in alphabetical order, along with the addresses of the associated Web sites.

ADAPT (<http://bmsr.usc.edu/Software/Adapt/adptmenu.html>): A suite of programs for fitting pharmacokinetic and pharmacodynamic models. The Web site also includes user-supplied functions for fitting complex models. Work on a version of the package that will allow population or hierarchical modeling is underway.

MCSIM (http://toxi.ineris.fr/activites/toxicologie_quantitative/mcsim/mcsim.php): A package that allows one to fit one's own statistical or simulation models and carry out Bayesian inference via Markov chain Monte Carlo simulations. It is useful for physiologically-based pharmacokinetic (or toxicokinetic) modeling.

PKBUGS (http://www.med.ic.ac.uk/divisions/60/pkbugs_web/home.html): An add-on program for carrying out fully Bayesian pharmacokinetic analyses within WinBUGS (<http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml>). One can modify the code generated by the program for problem-specific analyses.

ACKNOWLEDGMENTS

We thank the Children's Oncology Group and Dr. Lisa Bomgaars for permission to discuss the infant methotrexate study. This work was partially supported by grants CA75984, and GM61393.

REFERENCES

1. Johansen MJ, Newman RA, Madden T. The use of microdialysis in pharmacokinetics and pharmacodynamics. *Pharmacotherapy* 1997;17(3):464–481.
2. Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R, Hewett M, Lin Z, Liu Y, Liu S, Oliver DE, Rubin DL, Shafa F, Stuart JM, Altman RB. Integrating genotype and phenotype information: an overview of the PharmGKB Project. *Pharmacogenomics J.* 2001;1:167–170.
3. Licinio J, Wong M-L, eds. *Pharmacogenomics: The Search for Individualized Therapies*. Weinheim: Wiley-VCH; 2002.
4. Gibaldi M, Perrier D. *Pharmacokinetics*. 2nd ed. New York: Marcel Dekker, Inc.; 1982.
5. Sheiner LB, Rosenberg B, Melmon KL. Modelling of individual pharmacokinetics for computer-aided drug dosage. *Comput. Biomed. Res.* 1972;5(5):411–459.
6. Sheiner LB, Rosenberg B, Marathe VV. Estimation of population characteristics of pharmacokinetic parameters from routine clinical data. *J Pharmacokinet. Biopharm.* 1977;5(5):445–479.

7. Sheiner LB, Beal S, Rosenberg B, Marathe VV. Forecasting individual pharmacokinetics. *Clin. Pharmacol. Ther.* 1979;26(3):294–305.
8. Racine-Poon A, Smith AFM. Population models. In: Berry DA, ed., *Statistical Methodology in the Pharmaceutical Sciences*. New York: Marcel Dekker, Inc.; 1990:139–162.
9. Davidian M, Giltinan DM. *Nonlinear Models for Repeated Measurement Data*. London: Chapman and Hall; 1995.
10. Schwartz GJ, Brion LP, Spitzer A. The use of plasma creatinine concentration for estimating glomerular filtration rate in infants, children, and adolescents. *Pediatr. Clin. North. Am.* 1987;34(3):571–590.
11. Schwartz GJ, Feld LG, Langford DJ. A simple estimate of glomerular filtration rate in full-term infants during the first year of life. *J. Pediatr.* 1984;104(6):849–854.
12. Lunn DJ, Wakefield J, Thomas A, Best N, Spiegelhalter D. PKBugs User Guide, Version 1.1. In London, U.K.: Department of Epidemiology & Public Health, Imperial College of Science, Technology and Medicine; 1999.
13. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS, A Bayesian Modelling Framework: Concepts, Structure, and Extensibility. *Stat. and Computing* 2000;10:325–337.
14. Wakefield J. The Bayesian analysis of population pharmacokinetic models. *J. Am. Statist. Assoc.* 1996;91(433):62–75.
15. Mallet A. A maximum likelihood estimation method for random coefficient regression models. *Biometrika* 1986;73(3):645–656.
16. Mallet A, Mentré F, Steimer J-L, Lokiec F. Nonparametric maximum likelihood estimation for population pharmacokinetics, with application to cyclosporine. *J. Pharmacokin. Biopharm.* 1988;16(3):311–327.
17. Schumitzky A. Nonparametric EM algorithms for estimating prior distributions. *Appl. Math. Comput.* 1991;45:141–157.
18. Schumitzky A. The nonparametric maximum likelihood approach to pharmacokinetic population analysis. In: *Western Simulation Multiconference: Simulation in Health Care*; 1993; San Diego: Society for Computer Simulation; 1993. 95–100.
19. Laird N. Nonparametric maximum likelihood estimation of a mixing distribution. *J. Am. Stat. Assoc.* 1978;73:805–811.
20. Lindsay BG. The geometry of mixture likelihoods: a general theory. *Ann. Statist.* 1983;11:86–94.
21. Mallet A, Mentré F, Gilles J, Kelman AW, Thomson AH, Bryson SM, Whiting B. Handling covariates in population pharmacokinetics, with an application to Gentamicin. *Biomed. Meas. Inform. Contr.* 1988;2:138–146.
22. Gallant AR, Nychka DW. Semi-nonparametric maximum likelihood estimation. *Econometrica* 1987;55(2):363–390.
23. Davidian M, Gallant AR. The nonlinear mixed effects model with a smooth random effects density. *Biometrika* 1993;80(3):475–488.
24. Wakefield J, Walker S. Bayesian nonparametric population models: formulation and comparison with likelihood approaches. *J. Pharmacokin. Biopharm.* 1997;25(2):235–253.
25. Ferguson TS. A Bayesian analysis of some nonparametric problems. *Ann. Statist.* 1973;1:209–230.
26. West M, Müller P, Escobar MD. Hierarchical priors and mixture models, with applications in regression and density estimation. In: Smith AFM, Freeman P, eds. *Aspects of Uncertainty: A Tribute to D V Lindley*. New York: John Wiley & Sons; 1994:363–386.

27. Escobar MD, West M. Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.* 1995;90(430):577–588.
28. Müller P, Rosner G. Semiparametric PK/PD models. In: *Practical Nonparametric and Semiparametric Bayesian Statistics*. New York: Springer-Verlag; 1998:323–337.
29. Antoniak CE. Mixtures of Dirichlet processes with applications to nonparametric problems. *Ann. Statist.* 1974;2:1152–1174.
30. Rosner GL, Müller P. Bayesian population pharmacokinetic and pharmacodynamic analyses using mixture models. *J. Pharmacokin. Biopharm.* 1997;25(2):209–233.
31. Müller P, Rosner GL. A Bayesian population model with hierarchical mixture priors applied to blood count data. *J. of the Am. Stat. Assoc.* 1997;92:1279–1292.
32. Karlsson MO, Port RE, Ratain MJ, Sheiner LB. A population model for the leukopenic effect of etoposide. *Clin. Pharmacol. Ther.* 1995;57:325–334.
33. Unadkat JD, Bartha F, Sheiner LB. Simultaneous modeling of pharmacokinetics and pharmacodynamics with nonparametric kinetic and dynamic models. *Clin. Pharmacol. Ther.* 1986;40(1):86–93.
34. Verotta D, Sheiner LB. Simultaneous modeling of pharmacokinetics and pharmacodynamics: an improved algorithm. *Computer Appl. in the Biosciences* 1987;3(4):345–349.
35. Sheiner LB, Stanski DR, Vozeh S, Miller RD, Ham J. Simultaneous modeling of pharmacokinetics and pharmacodynamics: application to *d*-tubocurarine. *Clin. Pharmacol. Ther.* 1979;25:358–371.
36. Minami H, Sasaki Y, Saijo N, Ohtsu T, Fujii H, Igarashi T, Itoh K. Indirect-response model for the time course of leukopenia with anticancer drugs. *Clin. Pharmacol. Ther.* 1998;64(5):511–521.
37. Minami H, Sasaki Y, Watanabe T, Ogawa M. Pharmacodynamic modeling of the entire time course of leukopenia after a 3-hour infusion of paclitaxel. *Jpn. J. Cancer Res.* 2001;92(2):231–238.
38. Lopes HF, Müller P, Rosner GL. Bayesian meta-analysis for longitudinal data models using multivariate mixture priors. *Biometrics* 2003;59:66–75.
39. De Iorio M, Müller P, Rosner GL, MacEachern SN. An ANOVA model for dependent random measures. *J. Am. Stat. Assoc.* 2004;99:205–215.
40. Jelliffe RW. Clinical applications of pharmacokinetics and control theory: planning, monitoring, and adjusting dosage regimens of aminoglycosides, lidocaine, digitoxic, and digoxin. In: Maronde R, ed. *Selected Topics in Clinical Pharmacology*. New York: Springer Verlag; 1986:26–82.
41. Jelliffe RW, Iglesias T, Hurst AK, Foo KA, Rodriguez J. Individualising gentamicin dosage regimens: a comparative review of selected models, data fitting methods, and monitoring strategies. *Clin. Pharmacokin. Concepts* 1991;21(6):461–478.
42. Jelliffe RW, Schumitzky A, Guilder MV, Liu M, Hu L, Maire P, Gomis P, Barbaut X, Tahani B. Individualizing drug dosage regimens: roles of population pharmacokinetic and dynamic models, Bayesian fitting, and adaptive control. *Ther. Drug Monit.* 1993;15(5):380–393.
43. Collins JM. Pharmacokinetics and Clinical Monitoring. In: Chabner BW, Collins JM, eds. *Cancer Chemotherapy: Principles and Practice*. Philadelphia: Lippincott; 1990:16–31.
44. Ratain MJ, Mick R, Schilsky RL, Vogelzang NJ, Berezin F. Pharmacologically based dosing of etoposide: a means of safely increasing dose intensity. *J. Clin. Oncol.* 1991;9(8):1480–1486.
45. Chabner BA, Collins JM. *Cancer Chemotherapy: Principles and Practice*. Philadelphia: J.B. Lippincott Company; 1990.

46. Chaloner K, Verdinelli I. Bayesian experimental design: a review. *Statist. Sci.* 1995;10(3):273–304.
47. Merlé Y, Mentré F, Mallet A, Aurengo AH. Designing an optimal experiment for Bayesian estimation: application to the kinetics of iodine thyroid uptake. *Stat. in Med.* 1994;13(2):185–196.
48. Merlé Y, Mentré F. Bayesian design criteria: computation, comparison, and application to a pharmacokinetic and a pharmacodynamic model. *J. Pharmacokin. Biopharm.* 1995;23(1):101–125.
49. D’Argenio DZ. Incorporating prior parameter uncertainty in the design of sampling schedules for pharmacokinetic parameter estimation experiments. *Math. Biosci.* 1990;99(1):105–118.
50. Stroud JR, Müller P, Rosner GL. Optimal sampling times in population pharmacokinetic studies. *Appl. Stat.* 2001;50(3):345–359.
51. Gianni L, Kearns CM, Giani A, Capri G, Viganó L, Locatelli A, Bonadonna G, Egorin MJ. Nonlinear pharmacokinetics and metabolism of paclitaxel and its pharmacokinetic/pharmacodynamic relationships in humans. *J. Clin. Oncol.* 1995;13(1):180–190.
52. Gamerman D. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Boca Raton: CRC Press; 1998.
53. Müller P. Simulation-based optimal design. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM, eds. *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting*. Oxford: Oxford University Press; 1999:459–474.
54. Müller P, Parmigiani G. Optimal design via curve fitting of Monte Carlo experiments. *J. Am. Stat. Assoc.* 1995;90:1322–1330.

5 Practical Implementation of the Continual Reassessment Method

Naoki Ishizuka and Satoshi Morita

CONTENTS

5.1	Introduction	97
5.2	Example of a Phase I Trial.....	98
5.3	Elicitation of the Dose-Response Model and Prior Distribution.....	99
5.4	Dose Allocation	102
5.5	Sample Size and the Stopping Rule.....	105
5.6	Conclusion	108
	References	109
Appendix 5.1	SAS Codes for Logistic Dose-Response Model	110
Appendix 5.2	SAS Codes for Density Function of MTD	111
Appendix 5.3	SAS Codes for CRM Calculation and Density Function of DLT Occurrence Rate	113

5.1 INTRODUCTION

In this chapter we discuss the continual reassessment method (CRM) for conducting phase I trials.¹ In particular, we emphasize the use of graphical presentations for the application of a Bayesian approach.² For many years, phase I trials of anticancer drugs have not involved statisticians because the conventional design with three-patient cohorts is still widely used.^{3,4} A first step is for statisticians to explain the differences between conventional and CRM designs to clinicians, emphasizing the potential benefits of the CRM. This first challenge is difficult because the conventional design is very simple to understand, whereas the CRM is more complicated. Clinicians may fear that what they now do routinely under the three-patient cohort method will no longer be allowed by the CRM. Many clinicians either do not understand the definition of maximum tolerated dose (MTD) or believe that it automatically corresponds to 33% of the dose limiting toxicity (DLT) occurrence rate. Clinicians may be unfamiliar with ongoing collaboration with a statistician throughout the trial and with the information a clinician should provide to aid in decisions to escalate or deescalate dose levels. In order to overcome these issues, a statistician

needs tools to communicate with nonstatisticians. Graphical presentations are one such powerful tool that can be used to elicit prior assumptions and to convey the posterior results from the CRM.

In Section 5.2, an example of a phase I trial is described to illustrate key concepts. Section 5.3 discusses how to form initial settings of a dose-response model and the prior distribution. Section 5.4 discusses how to allocate dosage levels. The use of graphical presentations is emphasized in each of these sections. Following a discussion of sample sizes and stopping rules in Section 5.5, Section 5.6 discusses theoretical issues to be explored and provides concluding remarks. SAS codes for all examples are listed in the appendix to this chapter.

5.2 EXAMPLE OF A PHASE I TRIAL

For illustration, we describe a phase I trial in leukemia. The target of the DLT occurrence rate is determined solely by the balance of the expected efficacy and the type of toxicity. Some patients who relapse after allogeneic hematopoietic stem cell transplantation for adult leukemia would be expected to show graft-versus-host disease (GVHD) because it is assumed to have both graft-versus-leukemia (GVL) and graft-versus-tumor (GVT) effects.⁵⁻⁸ To induce GVHD, alpha-interferon is an alternative to chemotherapy following transplantation. In this situation, the target toxicity occurrence rate would be much higher, for example, near 50%, compared to the more common values of 33% or less for other types of cancer. This is because DLT includes Grade 3 or Grade 4 of GVHD and other severe adverse events that are common for the other types of cancer. However, the moderate GVHD at the MTD is believed not only to be toxic but also to be able to cure the disease. The ultimate goal is to choose the dose level that cures the disease and increase the number of the patients who will survive.

A hypothetical trial is assumed as follows:

1. Target DLT occurrence rate is 50%.
2. Four dose levels of alpha-interferon are to be tested. The prior estimates of the proportions of patients experiencing toxicities at each of the dose levels are given in the following table, and are used to construct the dose-response model.

Dose Level	1	2	3	4
Prior estimate	20%	35%	55%	80%

3. Estimated probability at the allocated dose level, which is higher than 75% of the DLT occurrence rate, must be less than 0.1.
4. Starting dose level is the lowest.
5. No jumping of dose levels over more than one dose level is allowed in dose escalation.
6. Dose allocation is every two patients for the first 6 patients.
7. Feasible sample size is at most 30 patients.

To distinguish the notation of the DLT occurrence rate and the probability, the DLT occurrence rate is described as a percentage, while any value of probability is not included in this chapter.

5.3 ELICITATION OF THE DOSE-RESPONSE MODEL AND PRIOR DISTRIBUTION

We follow the notation of the CRM in the manner of O'Quigley, as described in the CRM chapter by O'Quigley in this volume. As O'Quigley et al.¹ illustrated, in the typical CRM model, the dose relationship is modeled by a one-parameter distribution, $\Pr(Y_j = 1|x_j) = \psi(x_j, a)$ where Y_j is a response of 0 or 1 for the j th patient, where 1 denotes toxicity and 0 otherwise, x_j is a coding for dose for the j th patient, and a is a model parameter. If one assumes a logistic function for, $\psi(x_j, a)$, O'Quigley and Chevret⁹ suggest a fixed intercept model of

$$\Pr(Y_j = 1|x_j) = \psi(x_j, a) = \frac{\exp(3 + ax_j)}{1 + \exp(3 + ax_j)},$$

where $\text{logit } \Pr(Y_j = 1|x_j) = 3 + a \cdot x_j$.

As a typical prior, the following gamma distribution is used:

$$g(a) = \frac{1}{\Gamma(\lambda)} \left(\frac{1}{\lambda} \right)^\lambda a^{\lambda-1} \exp \left\{ - \left(\frac{a}{\lambda} \right) \right\},$$

where Γ is the gamma function and λ is the shape parameter. The scale parameter, $1/\lambda$, is defined so that $E(a) = 1$. If $\lambda = 1$, then, $g(a) = \exp(-a)$, the unit exponential function. The resulting dose response can be graphed with Bayesian confidence intervals. To illustrate the influence of the intercept, Figure 5.1 shows two figures for intercepts -3 and 3 under a unit exponential prior distribution for the parameter a . The vertical axis is DLT occurrence rate, and the horizontal axis is x , which in these figures is dose. Higher values of x correspond to higher doses in both figures. Even though x is positive in the left panels and negative in the right panels, the shape of the logistic curve of the dose-response and the vertical axis are exactly the same. The parameterizations of dose are very arbitrary. Thus, researchers can choose dose level x that they believe corresponds to the prior estimates of the DLT occurrence rate like those in the above table. The regions of narrower and wider confidence intervals are obvious in these graphs. If clinicians are more confident about lower doses than higher doses, it would be suitable to choose -3 as the intercept.

This is also reflected in the prior density function of MTD in Figure 5.2. If one assumes the logistic model and that the target of DLT occurrence rate is 50%, then the random variable MTD is transformed from the parameter a as:

$$\text{logit}(0.5) = \text{intercept} + a \cdot \text{MTD},$$

$$\text{MTD} = -\text{intercept}/a.$$

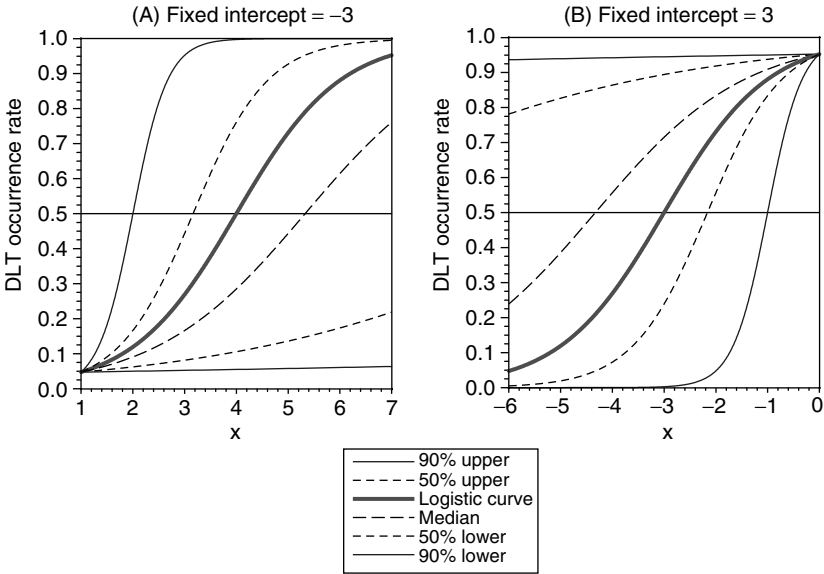


FIGURE 5.1 Assumed logistic dose response model and confidence intervals: (A) fixed intercept = -3 ; (B) fixed intercept = 3 .

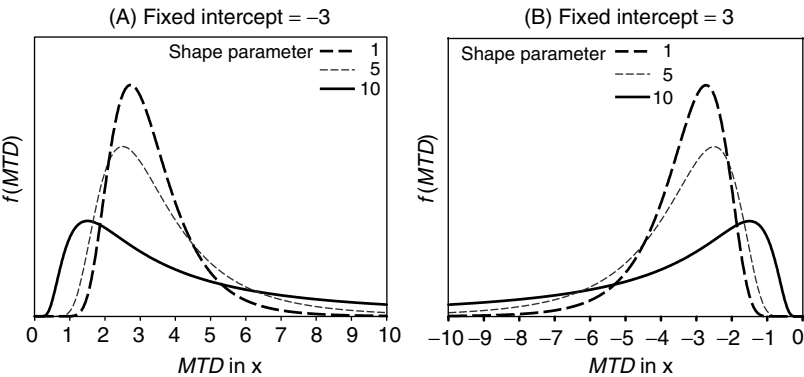


FIGURE 5.2 The prior density function of MTD and the shape parameter of gamma distribution for parameter a where $E(a) = 1$: (A) fixed intercept = -3 ; (B) fixed intercept = 3 .

Then, the density function $f(MTD)$ of MTD is derived as:

$$g(a)da = f(MTD)dMTD,$$
$$f(MTD) = g(a)\left|\frac{dMTD}{da}\right|^{-1}.$$

Careful interpretation of the density function of MTD is needed because the unit of MTD in the above equation is different from the original dose levels. Since dose level x is chosen as it corresponds to the prior estimates of the DLT occurrence rate. Figures 5.1 and 5.2 show that both the DLT occurrence rate and the distribution of the MTD are affected by the choice of intercept.

Another important figure is the prior density function of the DLT occurrence rate for each dose level, which is needed to determine the choice of the prior distribution of parameter a . The random variable representing the DLT occurrence rate $R(x_j)$ is also transformed by the random variable of parameter a for a single dose level of interest. The density function is retrieved in the same manner as for the MTD.

$$\text{logit}[R(x_j)] = \text{intercept} + a \cdot x_j$$

$$f(R(x_j)) = g(a) \left| \frac{dR(x_j)}{da} \right|^{-1}$$

Figure 5.3 shows the density function of the prior distribution of DLT occurrence rate for each dose level when the gamma distribution with shape = 1, 5, 10, where the scale parameter = 1/shape; i.e., mean = 1, is assumed as a prior distribution for parameter a . The assumed dose response is logistic with the fixed intercept of -3 .

The above distribution can be reflected by the chosen shape parameter to reflect how confident the investigators are in the prior parameters. If they had adequate information before the beginning of the trial, they would choose a larger value for the shape parameter. Restricting $E(a) = 1$ in the prior distribution implies the same

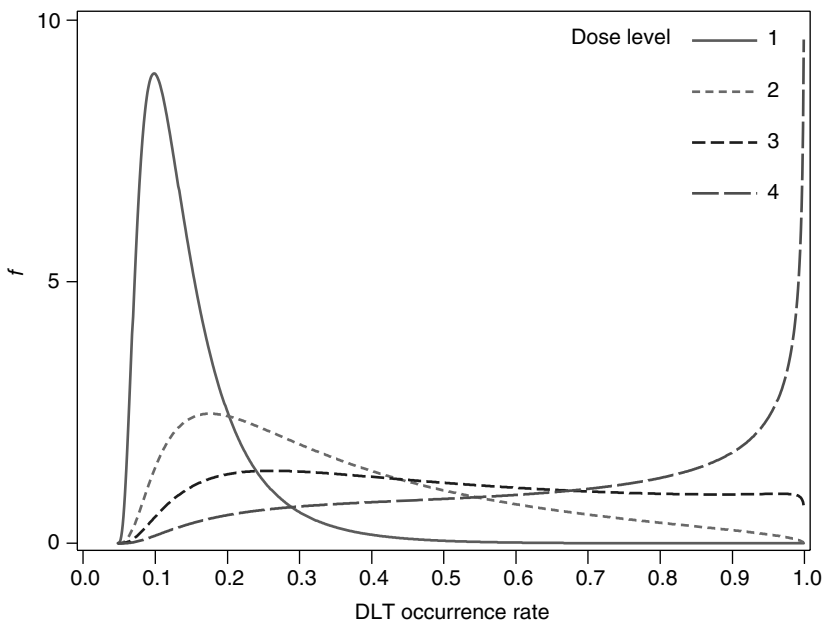


FIGURE 5.3 The prior density function of DLT occurrence rate with underlying gamma distribution for parameter a , where shape parameter = 5.

prior estimate for the DLT occurrence rate regardless of the magnitude of confidence. In general, if one were less confident, the prior distribution would be uniformly distributed. However, one's confidence is not monotonic across the dose levels because the choice of intercept concerning a logistic model affects the confidence in dose-response, as seen in [Figure 5.1](#). Notice also that the distribution of MTD is skewed; the right tail is flatter in [Figure 5.2](#). Thus, there is still a vague prior for higher than Level 1 if the shape parameter is 5. In contrast, if we use 3 as the fixed intercept concerning the logistic model, the retrieved prior density function of the DLT occurrence rate is distributed uniformly except for Level 4.

The relationship between the dose-response model, MTD, and the DLT occurrence rate for each dose level and parameter a , which is previously assumed, should be determined simultaneously. The prior distribution is somewhat arbitrary, as discussed by Gasparini and Eisele¹⁰ and Legedza and Ibrahim.¹¹ However, it is wise to try to elicit prior distributions not only for the model parameter of the dose-response model but also for the MTD or DLT occurrence rate at each dose level. As a matter of fact, our clinicians can determine the initial prior distribution when presented with these graphs, even those who are unfamiliar with the Bayesian approach.

5.4 DOSE ALLOCATION

There have been several modifications of the CRM in this decade; see Faries,¹² Korn et al.,¹³ Goodman et al.,¹⁴ Piantadosi et al.,¹⁵ and Ahn.¹⁶ The relevant question regarding dose escalation is whether the allocated dose level is near the target but safe enough to avoid an unacceptable toxicity level. To address the above question, the statistician must show

1. The distance between the target and the expected DLT occurrence rate $|\hat{R}(x_j) - \theta|$
2. The probability is more than a certain level of DLT occurrence rate; for example, 75%, i.e. $\Pr[R(x_j) > 75\%]$.

In the original CRM proposed by O'Quigley,¹ the expected DLT occurrence rate for each dose level is strictly considered. However, the approximation,

$$\hat{R}(x_j) \approx \psi(E[a], x_j)$$

where $E[a] = \int_0^\infty g(a)da$, is widely used because of its easy computation. It is valid if the dose response model is linear. Furthermore, this approximation does not work well when the shape parameter is large, as in [Table 5.1](#) because the approximation does not take into account the dispersion and shape parameter of the prior distribution of a . The expected DLT occurrence rate,

$$\hat{R}(x_j) = \int_0^\infty \psi(a, x_j) g(a)da,$$

is affected by the variance or confidence in the prior distribution. The expected value is conservative if a less confident or vague prior distribution is used. The term $\hat{R}(x_j)$

TABLE 5.1**Influence of Intercept and Shape Parameter of Gamma Distribution on the Mean DLT Occurrence Rate****Intercept = 3, DLT Occurrence Rate [0.00%–95.26%]**

Dose Level	Meta-Meter x	Prior Estimate of DLT Occurrence Rate			
		Approximation	No Approximation Shape Parameter		
			1	5	10
1	1.61371	20.00%	46.39%	30.91%	26.67%
2	2.38096	35.00%	52.21%	41.20%	38.60%
3	3.20067	55.00%	60.07%	55.20%	54.79%
4	4.38629	80.00%	75.39%	77.91%	78.84%

Intercept = −3, DLT Occurrence Rate [4.74%–100.00%]

Dose Level	Meta-Meter x	Prior Estimate of DLT Occurrence Rate			
		Approximation	No Approximation Shape Parameter		
			1	5	10
1	−1.61371	20.00%	24.61%	22.09%	21.59%
2	−2.79933	35.00%	35.10%	36.75%	36.27%
3	−3.61904	55.00%	44.03%	52.02%	53.50%
4	−4.38629	80.00%	53.61%	69.09%	73.33%

should be calculated by numerical integration and used to show the distance from the target.

Even though the dose might be close to the target in expectation, it could never guarantee safety, which can be specified in terms of the probability $\Pr[R(x_j) > 0.75\%]$ required in this case to be less than 0.1. This probability is calculated as follows:

$$\Pr[R(x_j) > 0.75] = \int_{a^*}^{\infty} g(a) da,$$

$$\text{where } a^* = \frac{\log \text{it}(0.75) - \text{intercept}}{x_j}.$$

The following equation is used for the negative intercept:

$$\Pr[R(x_j) > 0.75] = \int_0^{a^*} g(a) da.$$

Suppose that our investigators agreed with the above assumption as follows:

- Logistic dose-response model with a fixed intercept of −3
- Gamma distribution for parameter a , where the shape parameter = 5

and that the study statistician has provided [Table 5.2](#) and [Figure 5.4](#). The CRM calculation for the first patient shows that Level 3 is the closest to the 50% target toxicity

TABLE 5.2
CRM Calculation for the First Patient

Dose Level	1	2	3	4
DLT occurrence rate				
Prior estimate: approximation by E(a)	20.0%	35.0%	55.0%	80.0%
Expected Value	22.1%	36.7%	52.0%	69.1%
Pr[DLT occurrence rate<75%]	0.995	0.930	0.765	0.500

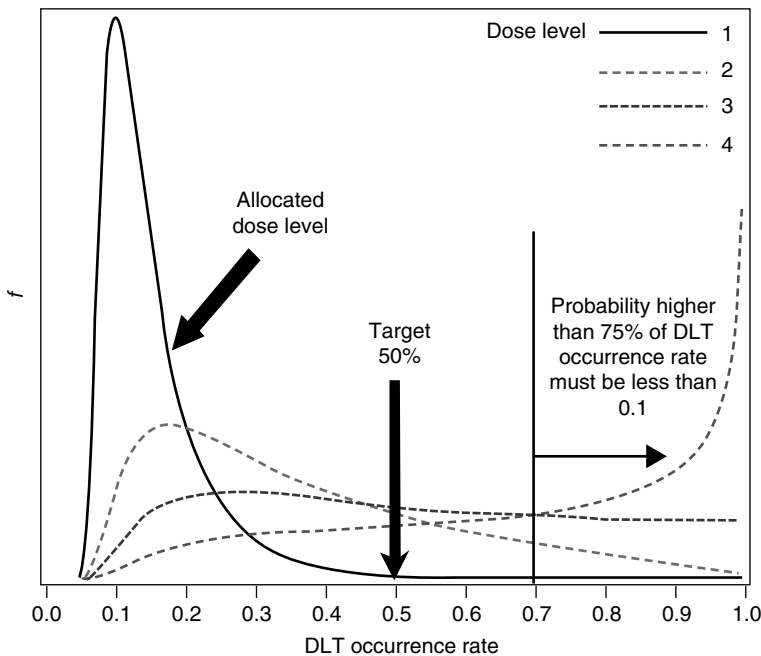


FIGURE 5.4 Allocation of dose level for the first patient and the prior density function of DLT occurrence rate.

and Level 2 is the probability for which more than 75% of DLT occurrence is estimated less than 0.1, as our criterion for controlling overdose. However, Level 1 is administered for the first patient based on the allocation rule mentioned in Section 5.2.

Some are concerned that it would take longer to accrue patients if the CRM is used because they think that they would have to wait to observe the response while the current patient is being treated. The three-patient cohort option, in which the dose allocation is determined for every three patients, would reduce such idle time. The dose allocation can be determined before the next response is observed. Furthermore, the same dose level will be allocated repeatedly if the number of treated patients in the trial becomes larger, indicating that the information is accumulating. To accelerate the study, it is useful to provide an allocation chart such as that shown in [Figure 5.5](#).

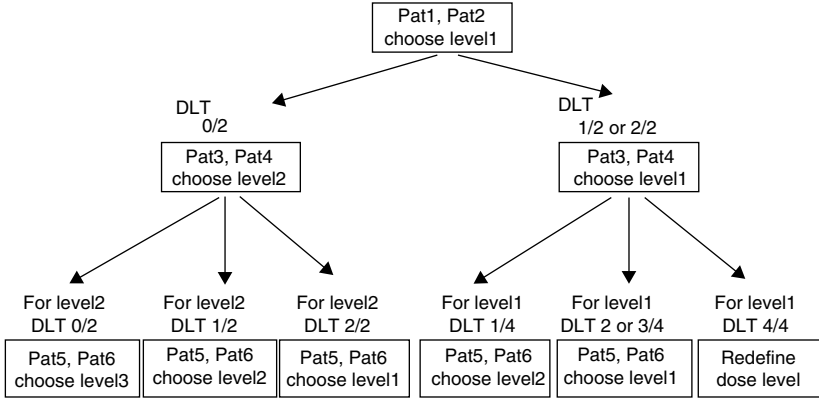


FIGURE 5.5 Dose allocation chart.

Even if the treated dose is not the allocated dose, tables of CRM calculations and allocation charts could be recalculated based on the observed responses to actual dose levels.

5.5 SAMPLE SIZE AND THE STOPPING RULE

In the example of the phase I study introduced in Section 5.2, the feasible sample size is at most 30 patients. This study has just begun, so we cannot discuss its results. However, we can discuss the issue of sample size and the stopping rule of CRM. Figure 5.6 shows a dose escalation history of a simulated trial with 30 patients. This simulated example selected Level 3 as the final recommendation for MTD, while Level 4 was administered to one patient (the 17th). Table 5.3 shows the CRM calculation for the 17th patient, who was treated after ten patients had already been treated at Level 3. DLTs were observed in 3 of those 10 patients. Table 5.3 also suggests that Level 3 was no longer the closest to 50%, as Level 4 was not only closer but also obtained a probability of more than 75% of DLT occurrence estimated as less than 0.1.

In addition to the tables, Figure 5.7 would be provided to physicians to assist in the decision concerning the next allocated dose level. The prior density function of the DLT occurrence rate would shift to the right if a DLT was observed, and it would shift to the left if a DLT was not observed. Once data have accumulated, the prior density function of the DLT occurrence rate becomes narrower, implying convergence, and it would not shift much further. As Heyd and Carlin¹⁷ discussed, one can terminate the trial if the 95% confidence intervals for prior distribution of MTD become sufficiently narrow. Ishizuka and Ohashi² suggested that one can terminate a trial if the prior density function of the DLT occurrence rate for each dose level is clearly separated from the others, even if the range of confidence intervals is less narrow. Figure 5.8 and Table 5.4 indicate that the results meet the criteria both for stopping the trial and the appropriateness of the selected dose.

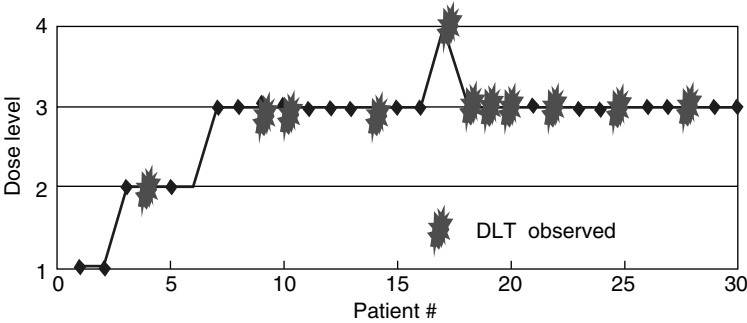


FIGURE 5.6 Dose escalation history for a simulated trial.

TABLE 5.3
CRM Calculation for the 17 Patient

Dose Level	1	2	3	4
DLT occurrence rate				
Prior estimate: approximation by E(a)	13.4%	20.9%	32.0%	52.0%
Expected Value	13.7%	21.7%	33.0%	51.7%
Pr[DLT occurrence rate<75%]	0.999	0.999	0.999	0.917

TABLE 5.4
CRM Calculation of the Final Result

Dose Level	1	2	3	4
DLT occurrence rate				
Prior estimate: approximation by E(a)	15.6%	25.7%	40.2%	63.8%
Expected Value	15.8%	26.0%	40.5%	63.1%
Pr[DLT occurrence rate<75%]	0.999	0.999	0.999	0.846

The fixed sample size has some advantages in conducting clinical trials. It allows the prediction of the required resources and the time required to complete the study. These are key issues in developing new treatments. As Green et al.⁴ noted, “Achieving a statistically accurate assessment of the MTD of a new agent would take hundreds of patients.” However, a single statistical test often is not used to determine the MTD in phase I trials because the MTD is required to be both accurate and precise. Thall et al.¹⁸ suggested fixed sample sizes, which means that no adoptive stopping rule is employed, and sizes should be large enough to be robust even when the initial prior is moderately different from the truth. Some simulations with fixed sample sizes should have been performed to guarantee robustness such that the final result will be close to the target DLT occurrence rate before the trial starts. Cheung and Chappell¹⁹ also proposed a technique to evaluate model sensitivity using a simulation to plan a phase I trial.

It would be inappropriate to have both a fixed sample size and a simple stopping rule; for example, to stop if six successive patients would be allocated to the same

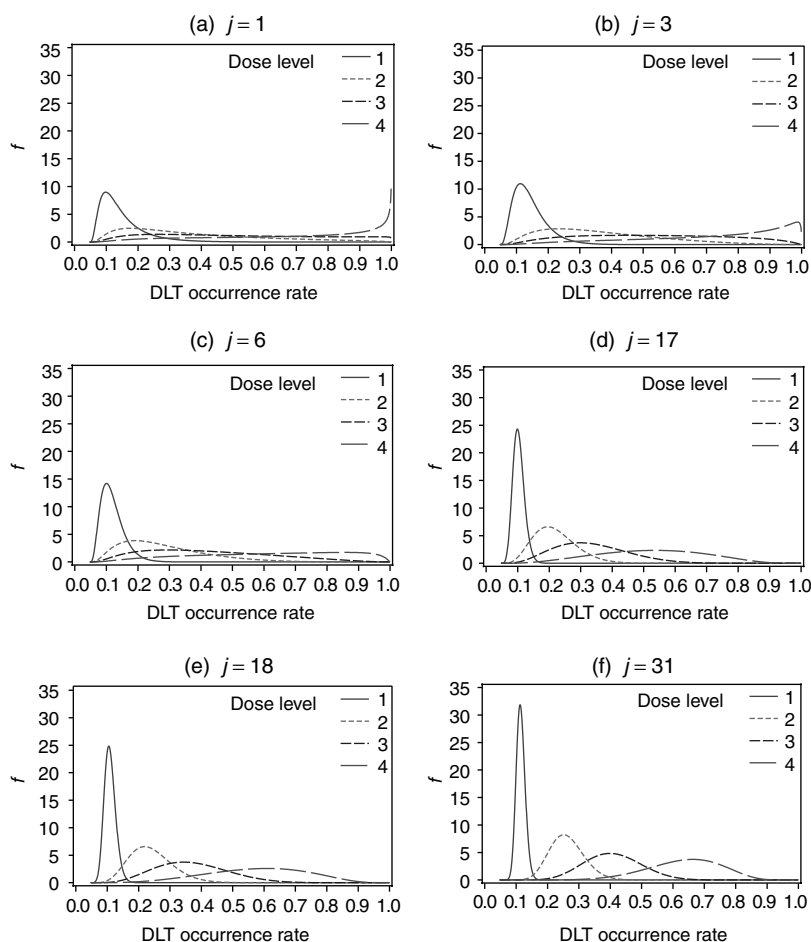


FIGURE 5.7 Updating the prior density function of DLT occurrence rate for the j th patient where they have the same scale for vertical and horizontal axis: (a) $j = 1$; (b) $j = 5$; (c) $j = 7$; (d) $j = 17$; (e) $j = 18$; (f) $j = 31$.

dose levels, as Goodman et al.¹⁴ discussed. O’Quigley and Reiner²⁰ suggested that it is possible to introduce a dynamic stopping rule other than the above approach, such as the use of 95% confidence of the prior distribution of MTD. To do so it is necessary to calculate the probability that the current dose level will be the final recommended dose at the end of the trial. As Figure 5.6 shows, Level 6 was repeatedly allocated from the 18th patient forward. Some would feel the trial should have been terminated at 25 patients. O’Quigley²¹ also presented this alternative, as the rule is not straightforward to implement. However, it does not require any additional calculations if one adopts a deterministic rule. One can implement the CRM calculation if the remaining patients are all extreme cases; for example, DLTs would be observed, or no DLT would be observed in any of them. If the recommended dose is

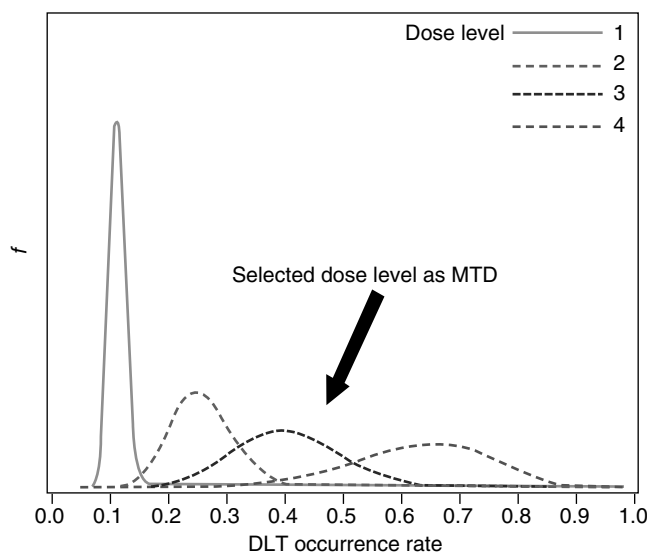


FIGURE 5.8 The posterior density function of DLT occurrence rate with 30 patients and selected dose level as MTD.

the same for both cases, one can conclude that the result has been achieved. In the simulated example of [Figure 5.6](#), Level 3 was chosen as the final recommendation after 28 patients. Regardless of the observation for the 29th and the 30th patient, the dose level allocated to the 31st patient is Level 3. Thus, we could have terminated the trial after 28 patients in this example. Graphical presentations are also useful for discussions of when to terminate a trial.

5.6 CONCLUSION

Graphical presentation is a powerful tool for the CRM, used to build the dose-response model and to elicit the prior distribution. The probability density function of the MTD and DLT occurrence rate should be provided for each dose level. The allocation chart is useful to prompt the trial. In addition to the CRM calculation, the monitoring of the prior distribution for DLT occurrence rate is essential to determine which dose level should be administered to the next patient. It is also useful in deciding when to terminate the trial even in the presence of a formal stopping rule.

Recent extensions of the CRM deal with an adjustment for patient heterogeneity; see Piantadosi et al.,¹⁵ Legedza and Ibrahim,¹¹ and O'Quigley et al.^{22,23} The introduction of additional parameters for patient covariates in the model makes the calculations and graphical presentation more complicated. It is sometimes unfeasible to perform numerical integration, but Markov chain Monte Carlo (MCMC) is an alternative in these situations.²⁴

REFERENCES

1. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics*. 1990;46(1):33–48.
2. Ishizuka N, Ohashi Y. The continual reassessment method and its applications: a Bayesian methodology for phase I cancer clinical trials. *Stat. Med.* 2001; 20(17–18): 2661–2681.
3. Piantadosi S. *Clinical Trials*. 1997 John Wiley and Sons, New York.
4. Green S, Benedetti J, Crowley J. *Clinical Trials in Oncology*. 2nd ed. (Interdisciplinary Statistics S.) 2003, Chapman & Hall/CRC.
5. Horowitz MM, et al. Graft-versus-leukemia reactions after bone marrow transplantation. *Blood* 1990;75:555–562.
6. Kolbs HJ, et al. Graft-versus-leukemia effect of donor lymphocyte transfusions in marrow grafted patients. *Blood* 1995;86:2041–2050.
7. Collins RH, et al. Hematologic relapse of chronic myelogenous leukemia following allogeneic bone marrow transplantation: apparent graft versus leukemia effect following abrupt discontinuation of immunosuppression. *Bone Marrow Transplant* 1992;10:391–395.
8. Porter DL, Connors JM, Van Deerlin VM, Duffy KM, McGarigle C, Saidman SL, et al. Graft-versus-tumor induction with donor leukocyte infusions as primary therapy for patients with malignancies. *J. Clin. Oncol.* 1999;17(4):1234.
9. 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(11):1647–1664.
10. Gasparini M, Eisele J. A curve-free method for phase I clinical trials. *Biometrics*. 2000;56(2):609–615.
11. Legedza AT, Ibrahim JG. Heterogeneity in phase I clinical trials: prior elicitation and computation using the continual reassessment method. *Stat. Med.* 2001;20(6):867–882.
12. Faries D. Practical modifications of the continual reassessment method for phase I cancer clinical trials. *J. Biopharm. Stat.* 1994;4(2):147–164.
13. Korn EL, Midthune D, Chen TT, Rubinstein LV, Christian MC, Simon RM. A comparison of two phase I trial designs. *Stat. Med.* 1994;13(18):1799–1806.
14. Goodman SN, Zahurak ML, Piantadosi S. Some practical improvements in the continual reassessment method for phase I studies. *Stat. Med.* 1995;14(11):1149–1161.
15. Piantadosi S, Liu G. Improved designs for dose escalation studies using pharmacokinetic measurements. *Stat. Med.* 1996;15(15):1605–1518.
16. Ahn C. An evaluation of phase I cancer clinical trial designs. *Stat. Med.* 1998;17(14):1537–1549.
17. Heyd JM, Carlin BP. Adaptive design improvements in the continual reassessment method for phase I studies. *Stat. Med.* 1999;18(11):1307–1321.
18. Thall PF, Lee JJ, Tseng CH, Estey EH. Accrual strategies for phase I trials with delayed patient outcome. *Stat. Med.* 1999;18(10):1155–1169.
19. Cheung YK, Chappell R. A simple technique to evaluate model sensitivity in the continual reassessment method. *Biometrics*. 2002;58(3):671–674.
20. O'Quigley J, Reiner E. A stopping rule for continual reassessment method. *Biometrika* 1998;85:741–748.
21. O'Quigley J. Continual reassessment designs with early termination. *Biostatistics*. 2002;3(1):87–99.
22. O'Quigley J, Shen LZ, Gamst A. Two-sample continual reassessment method. *J. Biopharm. Stat.* 1999;9(1):17–44.

23. O'Quigley J, Paoletti X. Continual reassessment method for ordered groups. *Biometrics*. 2003;59(2):430–440.
24. Gilks W.R., Richardson S., Spiegelhalter D.J. *Markov Chain Monte Carlo in Practice*. (Interdisciplinary Statistics S.) 1995, Chapman & Hall/CRC.

APPENDIX 5.1 SAS CODES FOR LOGISTIC DOSE-RESPONSE MODEL

```
axis1 order=0 to 1 by 0.1 label=(angle=270 h=2) major=
(h=1 w=2) minor=(n=3 h=0.5 w=1)w=5 offset=(0cm,0cm)
length=5inch;
axis2 label=(f=centxi h=2 'x') major=(h=1 w=2) minor=
(n=4 h=0.5 w=1)w=5 offset=(0cm,0cm) length=40pct;
legend1 position=(top right outside) value=(h=1.2 J=1
'90% Upper' '50% Upper' 'Logistic Curve' 'Median' '50%
Lower' '90% Lower')
mode=share across=1 shape=symbol(6,2) label=none ;
%macro gra_dra;

proc gplot;
plot p*x=1/vaxis=axis1 haxis=axis2 vref=0.5 legend=
legend1;
symbol1 v=none l=1 i=spline w=2 color=blue;
symbol2 v=none l=2 i=spline w=2 color=black;
symbol3 v=none l=1 i=spline w=5 color=red;
symbol4 v=none l=3 i=spline w=3 color=cyan;
symbol5 v=none l=2 i=spline w=2 color=black;
symbol6 v=none l=1 i=spline w=2 color=blue;
format p 3.1;
run;
quit;

%mend;
data work;
do l=1 to 6;
if l=1 then a=-log(0.95);
if l=2 then a=-log(0.75);
if l=3 then a=1;
if l=4 then a=-log(0.5);
if l=5 then a=-log(0.25);
if l=6 then a=-log(0.05);
a=a;
do x=0 to 5 by 0.01;
p=1-1/(1+exp(-3+a*x));
output;
end;
end;
```

```

label p='DLT occurrence rate';
run;
%gra_dra;

data work;
do l=1 to 6;
  if l=1 then a=-log(0.95);
  if l=2 then a=-log(0.75);
  if l=3 then a=-log(0.5);
  if l=4 then a=1;
  if l=5 then a=-log(0.25);
  if l=6 then a=-log(0.05);
  a=a;
do x=-5 to 0 by 0.01;
  p=1-1/(1+exp(3+a*x));
  output;
end;
end;
label p='DLT occurrence rate';
run;
legend1 position=(top right outside) value=(h=1.2 J=1
'90% Upper' '50% Upper' 'Median' 'Logistic Curve' '50%
Lower' '90% Lower')
mode=share across=1 shape=symbol(6,2) label=none;
%macro gra_drb;

proc gplot;
  plot p*x=1/vaxis=axis1 haxis=axis2 vref=0.5 leg-
end=legend1;
  symbol1 v=none l=1 i=spline w=2 color=blue;
  symbol2 v=none l=2 i=spline w=2 color=black;
  symbol3 v=none l=3 i=spline w=3 color=cyan;
  symbol4 v=none l=1 i=spline w=5 color=red;
  symbol5 v=none l=2 i=spline w=2 color=black;
  symbol6 v=none l=1 i=spline w=2 color=blue;
  format p 3.1;
run;
quit;

%mend;
%gra_drb;

```

APPENDIX 5.2 SAS CODES FOR DENSITY FUNCTION OF MTD

```

axis1 label=(h=2 f=centxi 'f' f=centx('f=centxi 'MTD'
f=centx ')) v=none w=5 minor=none major=none;

```

```

axis2 label=(h=2 f=centxi 'MTD in x') minor=none off-
set=(0cm,0cm) w=5;
legend1 position=(top right inside) value=(f=centx)
  mode=share across=1 shape=symbol(8,1) label=(f=centx
  h=1.5 'Shape Parameter');
title;

%macro gra_mtd;
proc gplot;
  plot f*x=shape
    / frame legend=legend1 vaxis=axis1 haxis=axis2;
  symbol1 i=spline v=none l=1 c=black w=3;
  symbol2 i=spline v=none l=1 c=red w=5;
  symbol3 i=spline v=none l=3 c=black w=3;
  symbol4 i=spline v=none l=5 c=black w=3;
run;
quit;
%mend;

data work;
  intercpt=-3;
do shape=1,5,10;
do x=0 to 10 by 0.01;
  a=-intercpt/x;
  prior=pdf('gamma',a,shape,1/shape);
  f=-prior*(intercpt/x**2);
  output;
end;
end;
run;

proc sort;
  by shape x;
run;

%gra_mtd;

data work;
  intercpt=3;
do shape=1,5,10;
do x=-10 to 0 by 0.01;
  a=-intercpt/x;
  prior=pdf('gamma',a,shape,1/shape);
  f=prior*(intercpt/x**2);
  output;
end;
end;
run;

proc sort;

```



```

by shape x;
run;
legend1 position=(top left inside) value=(f=centx)
mode=share across=1 shape=symbol(8,1) label=(f=centx
h=1.5 'Shape Parameter');
%gra_mtd;

```

APPENDIX 5.3 SAS CODES FOR CRM CALCULATION AND DENSITY FUNCTION OF DLT OCCURRENCE RATE

```

%macro crm(y10,y11,y20,y21,y30,y31,y40,y41);
proc iml;
  intercpt =-3;
  x1=1.61371; x2=2.38096; x3=3.20067; x4=4.38629;
  eps=1e-8; peak=0.1; scale=0.01;
  y10=&y10.; y20=&y20.; y30=&y30.; y40=&y40.;
  y11=&y11.; y21=&y21.; y31=&y31.; y41=&y41.;
start like(a)
  global (intercpt,x1,x2,x3,x4,y10,y11,y20,y21,y30,y31,y
  40,y41);
  v=exp(y11*(intercpt+a*x1)-(y10+y11)*log(1+exp(inter-
  cpt+a*x1))
  +y21*(intercpt+a*x2)-(y20+y21)*log(1+exp(intercpt+a
  *x2))
  +y31*(intercpt+a*x3)-(y30+y31)*log(1+exp(intercpt+a
  *x3))
  +y41*(intercpt+a*x4)-(y40+y41)*log(1+exp(intercpt+a
  *x4)));
  return(v);
finish;
start denom(a);
  pbb=pdf('gamma',a,5,0.2)*like(a);
  return(pbb);
finish;
start numeral(a);
  pbs1=a*pdf('gamma',a,5,0.2)*like(a);
  return(pbs1);
finish;
start numera2(a) global (intercpt,xi);
  pbs2=(1-1/(1+exp(intercpt+xi*a)))*pdf('gamma',a,
  5,0.2)*like(a);

```

```

    return(pbs2);
finish;
start invx(a);
    x=3/a;
    return(x);
finish;

range={0.p};
call quad(z0,"denom", range, eps, peak, scale);
call quad(z1,"numeral", range, eps, peak, scale);
mu=z1/z0;
xhat=invx(mu);
print z1 z0 mu xhat [format=f12.8];
print z0 [format=f14.12];

%do i=1 %to 4;
    xi=x&i.;
    call quad(z2&i., "numera2", range, eps, peak, scale);
    au=(log(3)+3)/xi;
    al={0};
    range2=al||au;
    call quad(z3&i., "denom", range2, eps, peak, scale);
%end;

phat=z21||z22||z23||z24;
phat=phat/z0;

p_low=z31||z32||z33||z34;
p_low=p_low/z0;

print phat [format=f12.8];

x=x1||x2||x3||x4;
pdash=1-1/(1+exp(-3+x#mu));
print pdash [format=f12.8];
print p_low [format=f12.8];

y0=y10||y20||y30||y40;
y1=y11||y21||y31||y41;

create z0 var{z0};
append from z0;
create y0 var{y10 y20 y30 y40};
append from y0;
create y1 var{y11 y21 y31 y41};
append from y1;

quit;
run;

```

```

data work;
  merge z0 y0 y1;
  x1=1.161371; x2=2.38096; x3=3.20067; x4=4.38629;
  /*x1=4; x2=6; x3=8; x4=11;*/intercpt=-3;
  array dose{4}x1-x4;
  do doselvl=1 to 4;
  do theta=0 to 1 by 0.005;
    a=(log(theta)-log(1-theta)+3)/dose{doselvl};
    like=exp(y11*(intercpt+a*x1)-(y10+y11)*log(1+
    exp(intercpt+a*x1))
    +y21*(intercpt+a*x2)-(y20+y21)*log(1+exp(intercpt+
    a*x2))
    +y31*(intercpt+a*x3)-(y30+y31)*log(1+exp(intercpt+
    a*x3))
    +y41*(intercpt+a*x4)-(y40+y41)*log(1+exp(intercpt+
    a*x4)));
    density=pdf('gamma',a,5,0.2)*like/z0;
    f=density/(theta*(1-theta)*dose{doselvl});
    if a>0 then output;
  end;
end;
label doselvl='Dose Level';
run;

proc sort data=work;
  by doselvl theta;
run;

data _null_;
  merge y0 y1;
  n=sum(y10,y11,y20,y21,y30,y31,y40,y41)+1;
  call symput("n",put(n,2.0));
run;

title1 h=2.0 "Prior Distribution of DLT occurrence rate
at each dose level for &n. pts";
axis1 value=none label=(h=2 f=centxi 'f') length=75pct
  w=2 major=none minor=none;
axis2 order=0 to 1 by 0.1 label=(h=2.5 'DLT occurrence
rate') value=(h=2) length=75pct minor=none w=2;
legend1 position=(top right inside) value=(h=2.0)
  mode=share across=1 shape=symbol(10,1) label=(f=cen
  tx h=2 'Dose Level');
proc gplot data=work;
  plot f*theta=doselvl
  / frame legend=legend1 vaxis=axis1 haxis=axis2;
  symbol1 i=join v=none l=1 w=1.5 c=cyan;

```

```
symbol2 i=join v=none l=2 w=1.5 c=magenta;  
symbol3 i=join v=none l=3 w=1.5 c=blue;  
symbol4 i=join v=none l=4 w=1.5 c=red;  
run;  
quit;  
  
%mend crm;  
*level1 level2 level3 level4;  
%crm(0,0, 0,0, 0,0, 0,0);*j=1;  
%crm(2,0, 1,1, 0,0, 0,0);*j=5;  
%crm(2,0, 3,1, 0,0, 0,0);*j=7;  
%crm(2,0, 3,1, 7,3, 0,0);*j=17;  
%crm(2,0, 3,1, 7,3, 0,1);*j=18;  
%crm(2,0, 3,1, 14,9, 0,1);*j=31;
```

Part II

Phase II Trials

6 Overview of Phase II Clinical Trials

Stephanie Green, Ph.D.

CONTENTS

6.1	Design.....	119
6.2	Analysis of Standard Phase II Designs	121
6.3	Other Phase II Designs.....	125
6.3.1	Multi-Arm Phase II Designs.....	125
6.3.2	Phase II Designs with Multiple Endpoints	125
6.3.3	Bayesian Phase II Designs	127
6.4	Discussion.....	128
	References	128

6.1 DESIGN

Standard phase II studies are used to screen new regimens for activity and to decide which ones should be tested further. To screen regimens efficiently, the decisions generally are based on single-arm studies using short-term endpoints, typically tumor response in cancer studies, in limited numbers of patients. The problem is formulated as a test of the null hypothesis $H_0: p = p_0$ vs. the alternative hypothesis $H_A: p = p_A$, where p is the probability of response, p_0 is the probability that if true would mean that the regimen was not worth studying further, and p_A is the probability that if true would mean it would be important to identify the regimen as active and to continue studying it. Typically, p_0 is a value at or somewhat below the historical probability of response to standard treatment for the same stage of disease, and p_A is typically somewhat above.

For ethical reasons studies of new agents usually are designed with two or more stages of accrual allowing early stopping due to inactivity of the agent. A variety of approaches to early stopping have been proposed. Although several of these include options for more than two stages, only the two-stage versions are discussed in this chapter. In typical clinical settings it is difficult to manage more than two stages. An early approach, due to Gehan,¹ suggested stopping if 0/ N responses were observed, where the probability of 0/ N was less than 0.05 under a specific alternative. Otherwise accrual was to be continued until the sample size was large enough for estimation at a specified level of precision.

In 1982 Fleming² proposed stopping when results are inconsistent either with H_0 or $H_{AA}: p = p'$, where H_0 is tested at level α and p' is the alternative for which the procedure has power $1-\alpha$. The bounds for stopping after the first stage of a two-stage design are the nearest integer to $N_1 p' - Z_{1-\alpha} \{N p' (1 - p')\}^{1/2}$ for concluding early that the regimen should not be tested further, and the nearest integer to $N_1 p_0 + Z_{1-\alpha} \{N p_0 (1 - p_0)\}^{1/2} + 1$ for concluding early that the regimen is promising, where N_1 is the first-stage sample size and N is the total after the second stage. At the second stage, H_0 is accepted or rejected according to the normal approximation for a single-stage design.

Since then other authors rather than proposing tests have proposed choosing stopping boundaries to minimize the expected number of patients required, subject to level and power specifications. Chang et al.³ proposed minimizing the average expected sample size under the null and alternative hypotheses. Simon,⁴ recognizing the ethical imperative of stopping when the agent is inactive, recommended stopping early only for unpromising results and minimizing the expected sample size under the null or, alternatively, minimizing the maximum sample size. A problem with these designs is that sample size has to be accrued exactly for the optimality properties to hold, so in practice they cannot be carried out faithfully in many settings. Particularly in multi-institution settings, studies cannot be closed after a specified number of patients have been accrued. It takes time to get out a closure notice, and during this time more patients will have been approached to enter the trial. Patients who have been asked and have agreed to participate in a trial should be allowed to do so, and this means there is a period of time during which institutions can continue registering patients even though the study is closing. Furthermore, some patients may be found to be ineligible after the study is closed. It is rare to end up with precisely the number of patients planned, making application of fixed designs problematic.

To address this problem, Green and Dahlberg⁵ proposed designs allowing for variable attained sample sizes. The approach is to accrue patients in two stages, to have level approximately 0.05 and power approximately 0.9, and to stop early if the agent appears unpromising. Specifically, the regimen is concluded unpromising and the trial is stopped early if the alternative $H_A: p = p_A$ is rejected in favor of $p < p_A$ at the 0.02 level after the first stage of accrual. The agent is concluded promising if $H_0: p = p_0$ is rejected in favor of $p > p_0$ at the 0.055 level after the second stage of accrual. The level 0.02 was chosen to balance the concern of treating the fewest possible patients with an inactive agent against the concern of rejecting an active agent due to treating a chance series of poor risk patients. Level 0.05 and power 0.9 are reasonable for solid tumors due to the modest percent of agents found to be active in this setting;⁶ less conservative values might be appropriate in more responsive diseases.

The design has the property that stopping at the first stage occurs when the estimate of the response probability is less than approximately p_0 , the true value that would mean the agent would not be of interest. At the second stage, the agent is concluded to warrant further study if the estimate of the response probability is greater than approximately $(p_A + p_0)/2$, which typically would be equal to or somewhat above the historical probability expected from other agents and a value at which one might be expected to be indifferent to the outcome of the trial. However, there are no optimality properties. Chen and Ng⁷ proposed a different approach to flexible design by optimizing with respect to expected sample size under p_0 across possible attained

sample sizes. They assumed a uniform distribution over sets of eight consecutive N_1 s; presumably if information is available on the actual distribution in a particular setting, then the approach could be used for a better optimization. To address the problem of temporary closure of studies, Herndon⁸ proposed an alternative approach that allows patient accrual to continue while results of the first stage are reviewed. Temporary closures are disruptive, so this approach might be reasonable for cases where accrual is relatively slow with respect to submission of information. If too rapid, the ethical aim of stopping early due to inactivity is lost.

Table 6.1 illustrates several of the design approaches mentioned above for level 0.05 and power 0.9 tests including Fleming designs, Simon minimax designs, Green and Dahlberg designs, and Chen and Ng optimal design sets. Powers and levels are reasonable for all approaches. Chen and Ng designs have the correct level on average, although individual realizations have levels up to 0.075 among the tabled designs. Of the four approaches, Green and Dahlberg designs are the most conservative with respect to early stopping for level 0.05 and power 0.9, whereas Chen and Ng designs are the least.

In another approach to phase II design, Storer⁹ suggested a procedure similar to two-sided testing instead of the standard one-sided test. In this approach, two hypothesis tests, one for H_0 and one for H_A are performed. The phase II is considered negative ($H_A: p \geq p_A$ is rejected) if the number of responses is sufficiently low, positive ($H_0: p \leq p_0$ is rejected) if sufficiently high, and equivocal if intermediate (neither hypothesis rejected). For a value p_m between p_0 and p_A , upper and lower rejection bounds are chosen such that the probability of concluding the trial is positive is less than γ when $p = p_m$, and the probability of concluding the trial is negative is also less than γ when $p = p_m$. The sample size and p_m are chosen to have adequate power to reject H_A under p_0 or H_0 under p_A . When $p_0 = 0.1$ and $p_A = 0.3$, an example of a Storer design is to test $p_m = 0.193$ with $\gamma = 0.33$ and power 0.8 under p_0 and p_A . For a two-stage design, N_1 , N , r_{L1} , r_{U1} , r_{L2} , and r_{U2} are 18, 29, 1, 6, 4, and 7, respectively, where N_1 is the sample size for the first stage, N is the total sample size, and r_{Li} and r_{Ui} are upper and lower rejection bounds for stage i , $i = 1, 2$. If the final result is equivocal (5 or 6 responses in 29 for this example), the conclusion is that other information is necessary to make the decision.

6.2 ANALYSIS OF STANDARD PHASE II DESIGNS

As noted in Storer, the hypothesis testing framework typically used in phase II studies is useful for developing designs and determining sample size. The resulting decision rules are not always meaningful, however, except as tied to hypothetical follow-up trials that in practice may or may not be done. Thus, it is important to present confidence intervals for phase II results that can be interpreted appropriately regardless of the nominal decision made at the end of the trial as to whether further study of the regimen is warranted.

The main analysis issue is estimation after a multistage trial, since the usual estimation procedures assuming a single-stage design are biased. Various approaches to generating confidence intervals have been proposed. These involve ordering the outcome space and inverting tail probabilities or test acceptance regions, as in

TABLE 6.1
Examples of Designs

	H_0 vs. H_A	N_1	a_1	b_1	N	a_2	b_2	Level (average and range for Chen)	Power (average and range for Chen)
Fleming	0.05 vs. 0.2	20	0	4	40	4	5	0.052	0.92
	0.1 vs. 0.3	20	2	6	35	6	7	0.053	0.92
	0.2 vs. 0.4	25	5	10	45	13	14	0.055	0.91
	0.3 vs. 0.5	30	9	16	55	22	23	0.042	0.91
Simon	0.05 vs. 0.2	29	1	—	38	4	5	0.039	0.90
	0.1 vs. 0.3	22	2	—	33	6	7	0.041	0.90
	0.2 vs. 0.4	24	5	—	45	13	14	0.048	0.90
	0.3 vs. 0.5	24	7	—	53	21	22	0.047	0.90
Green	0.05 vs. 0.2	20	0	—	40	4	5	0.047	0.92
	0.1 vs. 0.3	20	1	—	35	7	8	0.020	0.87
	0.2 vs. 0.4	25	4	—	45	13	14	0.052	0.91
	0.3 vs. 0.5	30	8	—	55	22	23	0.041	0.91
Chen	0.05 vs. 0.2	17–24	1	—	41–46	4	5	0.046	0.090
					47–48	5	6	0.022–0.069	0.845–0.946
	0.1 vs. 0.3	12–14	1	—	36–39	6	7	0.50	0.090
		15–19	2	—	40–43	7	8	0.029–0.075	0.848–0.938
		18–20	4	—	48	13	14	0.050	0.090
	0.2 vs. 0.4	21–24	5	—	49–51	14	15	0.034–0.073	0.868–0.937
		25	6	—	52–55	15	16		
		19–20	6	—	55	21	22	0.050	0.090
	0.3 vs. 0.5	21–23	7	—	56–58	22	23	0.035–0.064	0.872–0.929
		24–26	8	—	59–60	23	24		
					61–62	24	25		

N_1 is the sample size for the first stage of accrual, N is the total sample size after the second stage of accrual, a_i is the bound for accepting H_0 at stage i , and b_i is the bound for rejecting H_0 at stage i for $i = 1, 2$. Designs are listed for Fleming,² Simon,⁴ Green and Dahlberg;⁵ the optimal design set is listed for Chen and Ng.⁷

estimation following single-stage designs; however, with multistage designs the outcome space does not lend itself to any simple ordering.

Jennison and Turnbull¹⁰ order the outcome space by which boundary is reached, by the stage stopped at, and by the number of successes. Stopping at stage i is considered more extreme than stopping at stage $i + 1$ regardless of the number of successes. A value p is not in the $1-2\alpha$ confidence interval if the probability under p of the observed result or one more extreme according to this ordering is less than α in either direction. Chang and O’Brien¹¹ order the sample space instead based on the likelihood principle. For each p , the sample space for a two-stage design is ordered according to $L(x, N^*) = (x/N^*)^x(1-p)^{N^*-x}/p^x\{(N^*-x)/N^*\}^{N^*-x}$, where N^* is N_1 if the number of responses x can only be observed at the first stage and N if at the second. A value p is in the confidence interval if one half of the probability of the observed outcome plus the probability of a more extreme outcome according to this

ordering is α or less. The confidence set is not always strictly an interval, but the authors state that the effect of discontinuous points is negligible. Chang and O'Brien intervals are shorter than those of Jennison and Turnbull, although this in part would be because Jennison and Turnbull did not adjust for discreteness by assigning only one half of the probability of the observed value to the tail as Chang and O'Brien did. Duffy and Santner¹² recommend ordering the sample space by success percent and also develop intervals of shorter length than Jennison and Turnbull intervals.

Although they produce shorter intervals, Chang and O'Brien and Duffy and Santner approaches have the major disadvantage of requiring knowledge of the final sample size in order to calculate an interval for a study stopped at the first stage; as noted above this typically will be random. The Jennison and Turnbull approach can be used because it only requires knowledge up to the stopping time.

However, it is not entirely clear how important it is to adjust confidence intervals for the multistage nature of the design. From the point of view of appropriately reflecting the activity of the agent tested, the usual interval assuming a single-stage design may be sufficient. In this setting the length of the confidence interval is not of primary importance because sample sizes are small and all intervals are long. Similar to Storer's idea, it is assumed that if the confidence interval excludes p_0 , the regimen is considered active, and if it excludes p_A , the regimen is considered insufficiently active. If it excludes neither, results are equivocal. This seems reasonable whether or not continued testing is recommended for the better equivocal results.

For Green and Dahlberg designs, the differences between Jennison and Turnbull and unadjusted tail probabilities are 0 if the trial stops at the first stage,

$$-\sum_0^{a_1} \text{bin}(i, N_1, p) \sum_{x-i}^{N_2} \text{bin}(j, N-N_1, p) \text{ for the upper tail if stopped at the second stage,}$$

$$\text{and } + \sum_0^{a_1} \text{bin}(i, N_1, p) \sum_{x-i+1}^{N_2} \text{bin}(j, N-N_1, p) \text{ for the lower tail if stopped at the second}$$

stage, where a_1 is the stopping bound for accepting H_0 at the first stage. Both the upper and lower confidence bounds are shifted to the right for Jennison and Turnbull intervals. These therefore will more often appropriately exclude p_0 when p_A is true and inappropriately include p_A when p_0 is true compared to the unadjusted interval. However, the tail differences are generally small resulting in small differences in the intervals. Based on the normal approximation, the absolute value of the upper tail difference is less than approximately 0.003 when the lower bound of the unadjusted interval is p_0 , whereas the lower tail difference is constrained to be <0.02 for $p > p_A$ due to the early stopping rule. Generally the shift in a Jennison and Turnbull interval is noticeable only for small x at the second stage. As Rosner and Tsiatis¹³ note, such results, indicating activity in the first stage and no activity in the second, are unlikely, possibly suggesting the independent identically distributed assumption was incorrect.

For example, consider a common design for testing $H_0: p = 0.1$ vs $H_A: p = 0.3$: stop in favor of H_0 at the first stage if 0 or 1 responses are observed in 20 patients and otherwise continue to a total of 35. Of the 36 possible trial outcomes if planned sample sizes are achieved, the largest discrepancy in the 95% confidence intervals occurs if two responses are observed in the first stage and none in the second. For

this outcome, the Jennison and Turnbull 95% confidence interval is from 0.02 to 0.25, while the unadjusted interval is from 0.01 to 0.19. Although not identical, both intervals lead to the same conclusions: the alternative is ruled out.

For the Fleming and Green and Dahlberg designs listed in Table 6.1, Table 6.2 lists the probabilities that the 95% confidence intervals lie above p_0 (evidence the regimen is active), below p_A (evidence the agent has insufficient activity to pursue), or cover both p_0 and p_A (inconclusive). In no case are p_0 and p_A both excluded. Probabilities are calculated for $p = p_A$ and $p = p_0$, both for unadjusted and for Jennison and Turnbull adjusted intervals.

For the Green and Dahlberg designs considered, probabilities for the unadjusted and for the Jennison and Turnbull adjusted intervals are the same in most cases. The only discrepancy occurs for the 0.2 vs. 0.4 design when the final outcome is 11/45 responses. In this case the unadjusted interval is from 0.129 to 0.395, while the Jennison and Turnbull interval is from 0.131 to 0.402. There are more differences between adjusted and unadjusted probabilities for Fleming designs, the largest for ruling out p_A in the 0.2 vs. 0.4 and 0.1 vs. 0.3 designs. In these designs no second-stage Jennison and Turnbull interval excludes the alternative, making this probability unacceptably low under p_0 .

The examples presented suggest that adjusted confidence intervals do not necessarily result in more sensible intervals in phase II designs and in some cases are worse than not adjusting.

TABLE 6.2
Probabilities Under p_0 and p_A for Unadjusted and Jennison–Turnbull (J–T) Adjusted 95 % Confidence Intervals

		Probability 95% CI is above p_0 when $p =$		Probability p_A 95% CI is below p_A when $p =$		Probability 95% CI is includes p_0 and p_A when $p =$	
		p_0	p_A	p_0	p_A	p_0	p_A
0.05 vs. 0.2	Green J–T	0.014	0.836	0.704	0.017	0.282	0.147
	Unadjusted	0.014	0.836	0.704	0.017	0.282	0.147
	FlemingJ–T	0.024	0.854	0.704	0.017	0.272	0.129
	Unadjusted	0.024	0.854	0.704	0.017	0.272	0.129
0.1 vs. 0.3	Green J–T	0.020	0.866	0.747	0.014	0.233	0.120
	Unadjusted	0.020	0.866	0.747	0.014	0.233	0.120
	FlemingJ–T	0.025	0.866	0.392	0.008	0.583	0.126
	Unadjusted	0.025	0.866	0.515	0.011	0.460	0.123
0.2 vs. 0.4	Green J–T	0.025	0.856	0.742	0.016	0.233	0.128
	Unadjusted	0.025	0.856	0.833	0.027	0.142	0.117
	FlemingJ–T	0.023	0.802	0.421	0.009	0.556	0.189
	Unadjusted	0.034	0.862	0.654	0.022	0.312	0.116
0.3 vs. 0.5	Green J–T	0.022	0.859	0.822	0.020	0.156	0.121
	Unadjusted	0.022	0.859	0.822	0.020	0.156	0.121
	FlemingJ–T	0.025	0.860	0.778	0.025	0.197	0.115
	Unadjusted	0.025	0.860	0.837	0.030	0.138	0.110

6.3 OTHER PHASE II DESIGNS

6.3.1 MULTI-ARM PHASE II DESIGNS

Occasionally the aim of a phase II study is not to decide whether a particular regimen should be studied further but to decide which of several new regimens should be taken to the next phase of testing. In these cases selection designs are used, often formulated as follows: take on to further testing the treatment arm observed to be best by any amount, where the number of patients per arm is chosen to be large enough such that if one treatment is superior by Δ and the rest are equivalent the probability of choosing the superior treatment is p .

Simon et al.¹⁴ published sample sizes for selection designs with response endpoints, and Steinburg and Venzon¹⁵ proposed an approach to early selection in this setting. After the first stage of accrual for a two-arm trial, one of the treatments is chosen for further study if the number of responses is higher by at least a specified amount d_E , where d_E is chosen such that the probability of choosing an arm inferior by Δ is small. The procedure can be extended to three or more arms.

Liu et al.¹⁶ provide sample sizes for selection designs with survival endpoints. For survival the approach is to choose the arm with the smallest estimated β in a Cox model. Sample size is chosen so that if one treatment is superior with $\beta = -\ln(1 + \Delta)$ and the others have the same survival, then the superior treatment will be chosen with probability p .

Theoretically, selection designs are reasonable, but in reality the designs are not strictly followed. If response is poor in all arms, the conclusion should be to pursue none of the regimens, which is not an option for these designs. If a striking difference is observed, then the temptation is to bypass the confirmatory phase III trial. In a follow-up to the survival selection paper, Liu et al.¹⁷ noted that the probability of an observed $\beta < -\ln(1.7)$ which cancer investigators consider striking, is not negligible. With two to four arms, the probabilities are 0.07–0.08 when, in fact, there are no differences in the treatment arms.

6.3.2 PHASE II DESIGNS WITH MULTIPLE ENDPOINTS

The selected primary endpoint of a phase II trial is just one consideration in the decision to pursue a new regimen. If response is primary, secondary endpoints such as survival and toxicity must also be considered. For instance, a trial with a sufficient number of responses to be considered active may still not be of interest if too many patients experience life-threatening toxicity or if they all die quickly. On the other hand, a trial with an insufficient number of responses but a good toxicity profile and promising survival might still be considered for future trials.

Designs have been proposed to incorporate multiple endpoints explicitly into phase II studies. Bryant and Day¹⁸ proposed an extension of Simon's approach, identifying designs that minimize the expected accrual when the regimen is unacceptable either with respect to response or toxicity. Their designs are terminated at the first stage if either the number of responses is C_{R1} or less, or the number of patients without toxicity is C_{T1} or less, or both. The regimen is concluded useful if the number of patients with responses and the number without toxicity are greater than C_{R2} and C_{T2} ,

respectively, at the second stage. N_1 , N , C_{R1} , C_{T1} , C_{R2} , and C_{T2} are chosen such that the probability of recommending the regimen when the probability of no toxicity is acceptable ($p_T \geq p_{T1}$) but response is unacceptable ($p_R \leq p_{R0}$) is less than or equal to α_R , the probability of recommending the regimen when response is acceptable ($p_R \geq p_{R1}$) but toxicity is unacceptable ($p_T \leq p_{T0}$) is less than or equal to α_T , and the probability of recommending the regimen when both are acceptable is $1-\beta$ or better. The constraints are applied either uniformly over all possible correlations between toxicity and response or assuming independence of toxicity and response. Minimization is done subject to the constraints. For many practical situations, minimization assuming independence produces designs that perform reasonably well when the assumption is incorrect.

Conaway and Petroni¹⁹ proposed similar designs assuming a particular relationship between toxicity and response, an optimality criterion, and a fixed total sample size are all specified. Design constraints proposed include limiting the probability of recommending the regimen to α or less when both response and toxicity are unacceptable ($p_R \leq p_{R0}$ and $p_T \leq p_{T0}$) and to γ or less anywhere else in the null region ($p_R \leq p_{R0}$ or $p_T \leq p_{T0}$ but not both). The following year, Conaway and Petroni²⁰ proposed boundaries allowing for tradeoffs between toxicity and response. Instead of dividing the parameter space as in Figure 6.1(a), it is divided according to investigator specifications, as in Figure 6.1(b), allowing for fewer patients with no toxicity when the response probability is higher and the reverse.

Their proposed test accepts H_0 when a statistic $T(\mathbf{x})$ is less than c_1 at the first stage or is less than c_2 at the second, subject to maximum level α over the null region and power at least $1-\beta$ when $p_R = p_{R1}$ and $p_T = p_{T1}$ for an assumed value for the association between response and toxicity. The statistic $T(\mathbf{x})$ is $\sum p_{ij}^* \ln(p_{ij}^*/\hat{p}_{ij})$, where ij indexes the cells of the 2×2 response-toxicity table, $\hat{p}_{ij}$ are the usual probability estimates, and p_{ij}^* s are the values achieving $\inf_{H_0} \sum p_{ij} \ln(p_{ij}/\hat{p}_{ij})$. $T(\mathbf{x})$ can be interpreted in some sense as a distance from $\hat{p}$ to H_0 . Interim stopping bounds are chosen to satisfy optimality criteria. The authors' preference is minimization of the expected sample size under the null.

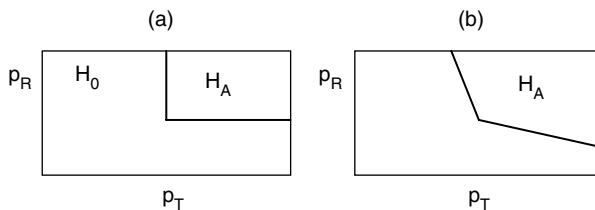


FIGURE 6.1 Division of parameter space for two approaches to bivariate phase II design. (a) An acceptable probability of response and an acceptable probability of no toxicity are each specified. (b) Acceptable probabilities are not fixed at one value for each but instead allow for trade-off between toxicity and response. p_R = probability of response and p_T = probability of acceptable toxicity.

Thall and Cheng²¹ proposed another approach to multi-endpoint design. Parameters of interest are $\Delta = (\Delta_R, \Delta_T)$, where $\Delta_R = g(p_{RI}) - g(p_{RO})$ is the difference between probability of response on experimental treatment (p_{RI}) and probability of historical response (p_{RO}), and $\Delta_T = g(p_{TI}) - g(p_{TO})$ is the difference between the probability of acceptable toxicity on experimental treatment (p_{TI}) and the probability of acceptable toxicity, historically (p_{TO}), after arcsine square root transformation. Target parameters (ξ_R, ξ_T) are identified and the alternative region is the set of all Δ s at least as desirable as the target, i.e., $\{\Delta : \Delta_R \geq \xi_R \text{ and } \Delta_T \geq \xi_T\}$. If multiple targets are identified, the alternative Ω_A is the convex hull of these regions. Trial outcome is $\hat{\Delta} = (g(\hat{p}_{RI}) - g(p_{RO}), g(\hat{p}_{TI}) - g(p_{TO}))$. The rejection region $R(x)$ is $\{(y-x, z-x) \mid (y,z) \in \Omega_A\}$, where the sample size n and a number x are chosen such that $\Pr\{\hat{\Delta} \in R(x) \mid \Delta = 0\} \leq \alpha$ and $\Pr\{\hat{\Delta} \in R(x) \mid \Delta = (\xi_R, \xi_T)\} \geq 1 - \beta$ for each target. The test is based on approximate bivariate normality of $\sqrt{4n}(\hat{\Delta} - \Delta)$. Interim stopping boundaries are based on optimality criteria.

There are a number of practical problems with these designs. As for other designs relying on optimality criteria, they generally cannot be done faithfully in realistic settings. Even when they can be carried out, defining toxicity as a single yes–no variable is problematic because typically several toxicities of various grades are of interest. Perhaps the most important issue is that of the response–toxicity trade-off. Any function specified is subjective and cannot be assumed to reflect the preferences of either investigators or patients in general.

6.3.3 BAYESIAN PHASE II DESIGNS

Bayesian approaches provide another formulation of phase II designs. As described in Estey and Thall,²² prior probability distributions are assigned to Π_H , the true historical probability of response, and to Π_E , the true probability of response of the regimen under study. The prior for Π_H is informative, whereas the prior for Π_E generally is not. After each specified interim analysis time the posterior distribution of Π_E , which also serves as the prior for the next stage of accrual, is calculated given the data. The distribution of Π_H is also updated if there is a randomized control arm, which the authors recommend. Accrual is stopped if the posterior probability that Π_E is at least Δ greater than Π_H is small. The maximum sample size is chosen such that the final posterior distribution for Π_E if accrual completes is sufficiently precise, and the regimen under study is considered worth further study if there is a reasonable probability that Π_E is Δ better than Π_H . As with any Bayesian designs, care must be taken that the a priori assumptions do not unduly influence the conclusion and that stopping criteria are sufficiently conservative.

The Bayesian framework has been used to address other phase II issues. For certain types of response endpoints, Cheung and Thall²³ addressed the problem of temporary study closure by proposing an adaptive Bayesian method. At each interim analysis time, an approximate posterior distribution is calculated using all of the event time data available including data from patients still on treatment for whom final endpoint determination is unknown. Nuisance parameters in the likelihood are replaced by consistent estimates. The design may reduce trial duration, but practical difficulties include the need for current follow-up and the numerous analyses.

Also of note, how to construct confidence intervals at the end of a study with a Bayesian design is not an issue. Intervals are generated from the final posterior distribution for Π_E with no special adjustment for number of looks at the data. Issues instead are the influence of the prior distribution and interpretation of the final interval.

6.4 DISCUSSION

Despite the precise formulation of decision rules, phase II trials are not as objective as we would like. The small sample sizes used cannot support decision-making based on all aspects of interest in a trial. Trials combining more than one aspect, such as toxicity and response, are fairly arbitrary with respect to the relative importance placed on each endpoint, including the 0 weight placed on the endpoints not included, and so are subject to about as much imprecision in interpretation as results of single endpoint trials. Furthermore, a phase II trial would rarely be considered on its own. By the time a regimen is taken to phase III testing, multiple phase II trials have been done and the outcomes of the various trials weighed and discussed. Perhaps statistical considerations in a phase II design are most useful in keeping investigators realistic about how limited such designs are.

For similar reasons, optimality considerations both with respect to design and confidence intervals are not particularly compelling in phase II trials. Sample sizes in the typical clinical setting are small and variable, making it more important to use procedures that work reasonably well across a variety of circumstances rather than optimally in one. Also, there are various characteristics that it would be useful to optimize; compromise is usually in order.

As a final practical note, choices of endpoints and null and alternative hypotheses in phase II trials should be reconsidered on an ongoing basis. As definitions and treatments change and patient populations used for testing new agents shift, old historical probabilities do not remain applicable. Thall and Wang in this volume discuss variability in single-arm trials and a possible approach to the problem of historical controls. As well, the primary endpoint used to assess efficacy should be reviewed. Other indicators of activity may be more suitable than tumor shrinkage, such as disease stability for cytostatic agents or specific biologic effects for targeted agents.

REFERENCES

1. Gehan EA. The determination of number of patients in a follow-up trial of a new chemotherapeutic agent. *J. Chronic. Dis.* 1961;13:346–353.
2. Fleming TR. One sample multiple testing procedures for phase II clinical trials. *Biometrics* 1982;38:143–151.
3. Chang MN, Therneau TM, Wieand HS, Cha SS. Designs for group sequential phase II clinical trials. *Biometrics* 1987;43:865–874.
4. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clin. Trials* 1989;10:1–10.
5. Green SJ and Dahlberg S. Planned vs attained design in phase II clinical trials. *Stat. Med.* 1992;11:853–862.

6. Simon R. How large should a phase II trial of a new drug be? *Cancer Treat. Rep.* 1987;71:1079–1085.
7. Chen T and Ng T-H. Optimal flexible designs in phase II clinical trials. *Stat. Med.* 1998;17:2301–2312.
8. Herndon J. A design alternative for two-stage, phase II, multicenter cancer clinical trials. *Controlled Clin. Trials* 1998; 19: 440–450.
9. Storer B. A class of phase II designs with three possible outcomes. *Biometrics* 1992;48:55–60.
10. Jennison C and Turnbull BW. Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials. *Technometrics* 1983;25:49–58.
11. Chang MN and O'Brien PC. Confidence intervals following group sequential tests. *Controlled Clin. Trials* 1986;7:18–26.
12. Duffy DE and Santner TJ. Confidence intervals for a binomial parameter based on multistage tests. *Biometrics* 1987;43:81–94.
13. Rosner G and Tsiatis A. Exact confidence intervals following a group sequential trial: a comparison of methods. *Biometrika* 1988;75:723–729.
14. Simon R, Wittes R, and Ellenberg S. Randomized phase II clinical trials. *Cancer Treat. Rep.* 1985;69:1375–1381.
15. Steinberg SM and Venzon DJ. Early selection in a randomized phase II clinical trial. *Stat. Med.* 2002;21:1711–1726.
16. Liu PY, Dahlberg S, and Crowley J. Selection designs for pilot studies based on survival endpoints. *Biometrics* 1993;49:391–398.
17. Liu PY, LeBlanc M, and Desai M. False positive rates of randomized phase II designs. *Controlled Clin. Trials* 1999;20:343–352.
18. Bryant J and Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics* 1995;51:1372–1383.
19. Conaway M and Petroni G. Bivariate sequential designs for phase II trials. *Biometrics* 1995;51:656–664.
20. Conaway M and Petroni G. Designs for phase II trials allowing for a trade-off between response and toxicity. *Biometrics* 1996;52:1375–1386.
21. Thall PF and Cheng S-C. Optimal two-stage designs for clinical trials, based on safety and efficacy. *Stat. Med.* 2001;20:1023–1032.
22. Estey EH and Thall PF. New designs for phase 2 clinical trials. *Blood* 2003;102:442–448.
23. Cheung YK and Thall PF. Monitoring the rates of composite events with censored data in phase II clinical trials. *Biometrics* 2002;58:89–97.

7 Designs Based on Toxicity and Response

*Gina R. Petroni, Ph.D. and
Mark R. Conaway, Ph.D.*

CONTENTS

7.1 Introduction	131
7.2 Designs for Response and Toxicity	132
7.3 Designs that Allow a Trade-Off between Response and Toxicity	137
7.4 Summary.....	141
References	142

7.1 INTRODUCTION

In principle, single-arm phase II trials evaluate whether a new regimen is sufficiently promising to warrant a comparison with the current standard of treatment. An agent is considered sufficiently promising based on the proportion of patients who respond, that is, experience some objective measure of disease improvement. The toxicity or adverse event profile of the new regimen, usually defined in terms of the proportion of patients experiencing severe adverse events, has been established in a previous phase I trial.

In practice, the separation between establishing the toxicity of a new agent in a phase I trial and establishing the response rate in a phase II trial is artificial. Most phase II trials are conducted not only to establish the response rate but also to gather additional information about the toxicity associated with the new agent. Conaway and Petroni¹ and Bryant and Day² cite several reasons why toxicity considerations are important for phase II trials:

1. Sample sizes in phase I trials. The number of patients in a phase I trial is small, and the toxicity profile of the new agent is estimated with little precision. As a result, there is a need to gather more information about toxicity rates before proceeding to a large comparative trial.
2. Ethical considerations. Most phase II trials are designed to terminate the study early if it does not appear that the new agent is sufficiently promising to warrant a comparative trial. These designs are meant to protect patients

from receiving substandard therapy. Patients should be protected also from receiving agents with excessive rates of toxicity, and consequently, phase II trials should be designed with the possibility of early termination of the study if an excessive number of toxicities is observed. This consideration is particularly important in studies of intensive chemotherapy regimens, where it is hypothesized that a more intensive therapy induces a greater chance of a response but also a greater chance of toxicity.

- 3. Patient differences. The characteristics of the patients enrolled in the previous phase I trials may be different than those of the patients to be enrolled in the phase II trial. For example, phase I trials often enroll patients for whom all standard therapies have failed. These patients are likely to have a greater extent of disease than patients who will be accrued to the phase II trial.

With these considerations, several proposals have been made for designing phase II trials that formally incorporate both response and toxicity endpoints. Conaway and Petroni¹ and Bryant and Day² propose methods that extend the two-stage designs of Simon.³ In each of these methods, a new agent is considered sufficiently promising if it exhibits both a response rate that is greater than that of the standard therapy and a toxicity rate that does not exceed that of the standard therapy. Conaway and Petroni⁴ consider a different criterion based on a trade-off between response and toxicity rates. In these designs a new agent with a greater toxicity rate might be considered sufficiently promising if it also has a much greater response rate than the standard therapy. Thall, Simon, and Estey^{5,6} propose a Bayesian method for monitoring response and toxicity that can also incorporate a trade-off between response and toxicity rates. Further modifications are given in Thall and Sung.⁷

7.2 DESIGNS FOR RESPONSE AND TOXICITY

Conaway and Petroni¹ and Bryant and Day² present multi-stage designs that formally monitor response and toxicity. As a motivation for the multi-stage designs, we first describe the methods for a fixed sample design using the notation in Conaway and Petroni.¹ In this setting, binary variables representing response and toxicity are observed in each of N patients. Table 7.1 summarizes the data in a 2×2 table where X_{ij} is the number of patients with response classification i , where $i = 1$ denotes

TABLE 7.1
Classification of Patients by Response and Toxicity

		Toxicity		Total
		Yes	No	
Response	Yes	X_{11}	X_{12}	X_R
	No	X_{21}	X_{22}	$N - X_R$
	Total	X_T	$N - X_T$	N

response and $i = 2$ denotes no response, and toxicity classification j , where $j = 1$ denotes toxicity and $j = 2$ denotes no toxicity. The observed number of responses is $X_R = X_{11} + X_{12}$, and the observed number of patients experiencing a severe toxicity is $X_T = X_{11} + X_{21}$. It is assumed that the cell counts in this table, $(X_{11}, X_{12}, X_{21}, X_{22})$, have a multinomial distribution with underlying probabilities $(p_{11}, p_{12}, p_{21}, p_{22})$. That is, in the population of patients to be treated with this new agent, a proportion p_{ij} would have response classification i and toxicity classification j as shown in Table 7.2. With this notation the probability of a response is $p_R = p_{11} + p_{12}$ and the probability of a toxicity is $p_T = p_{11} + p_{21}$.

The design is based on having sufficient power to test the null hypothesis that the new treatment is not sufficiently promising to warrant further study against the alternative hypothesis that the new agent is sufficiently promising to warrant a comparative trial. Conaway and Petroni¹ and Bryant and Day² interpret the term *sufficiently promising* to mean that the new treatment has a greater response rate than the standard and that the toxicity rate with the new treatment is no greater than that of the standard treatment. Defining p_{Ro} as the response rate with the standard treatment and p_{To} as the toxicity rate for the standard treatment, the null hypothesis can be written as

$$H_o: p_R \leq p_{Ro} \text{ or } p_T \geq p_{To}$$

$$H_a: p_R > p_{Ro} \text{ and } p_T < p_{To}$$

The null and alternative regions are displayed in Figure 7.1.

A statistic for testing H_o versus H_a is (X_R, X_T) , with a critical region of the form $C = \{(X_R, X_T): X_R \geq c_R \text{ and } X_T \leq c_T\}$. We reject the null hypothesis and declare the treatment sufficiently promising if we observe many responses and few toxicities. We do not reject the null hypothesis if we observe too few responses or too many toxicities. Conaway and Petroni¹ choose the sample size, N , and critical values (c_R, c_T) to constrain three error probabilities to be less than pre-specified levels α , γ , and $1 - \beta$, respectively. The three error probabilities are:

1. The probability of incorrectly declaring the treatment promising when the response and toxicity rates for the new therapy are the same as those of the standard therapy
2. The probability of incorrectly declaring the treatment promising when the response rate for the new therapy is no greater than that of the standard or

TABLE 7.2
Population Proportions for Response and Toxicity Classifications

		Toxicity		Total
		Yes	No	
Response	Yes	p_{11}	p_{12}	p_R
	No	p_{21}	p_{22}	$1 - p_R$
	Total	p_T	$1 - p_T$	1

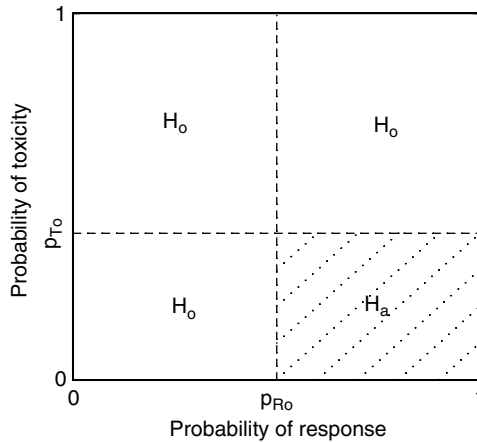


FIGURE 7.1 Null and alternative regions.

the toxicity rate for the new therapy is greater than that of the standard therapy.

3. The probability of declaring the treatment not promising at a particular point in the alternative region. The design should yield sufficient power to reject the null hypothesis for a specific response and toxicity rate, where the response rate is greater than that of the standard therapy and the toxicity rate is less than that of the standard therapy.

Mathematically, these error probabilities are:

$$P(X_R \geq c_R, X_T \leq c_T \mid p_R = p_{R_o}, p_T = p_{T_o}, \theta) \leq \alpha, \quad (7.1)$$

$$\sup_{p_R \leq p_{R_o} \text{ or } p_T \geq p_{T_o}} P(X_R \geq c_R, X_T \leq c_T \mid p_R, p_T, \theta) \leq \gamma, \quad (7.2)$$

$$P(X_R \geq c_R, X_T \leq c_T \mid p_R = p_{R_a}, p_T = p_{T_a}, \theta) \leq 1 - \beta \quad (7.3)$$

where these probabilities are computed for a pre-specified value of the odds ratio $\theta = (p_{11}p_{22})/(p_{12}p_{21})$ in Table 7.2. The point (p_{R_a}, p_{T_a}) is a pre-specified point in the alternative region, with $p_{R_a} > p_{R_o}$ and $p_{T_a} < p_{T_o}$.

Conaway and Petroni¹ compute the sample size and critical values by enumerating the distribution of (X_R, X_T) under particular values for (p_R, p_T, θ) . As an example, Conaway and Petroni¹ present a proposed phase II trial of high dose chemotherapy for patients with non-Hodgkin's lymphoma. Results from earlier studies for this patient population have indicated that standard therapy results in an estimated response rate of 50% with approximately 30% of patients experiencing life-threatening toxicities. In addition, previous results indicated that approximately 35% to 40% of the patients who experienced a complete response also experienced life-threatening toxicities. The odds ratio θ is determined by the assumed response rate,

toxicity rate, and the conditional probability of experiencing a life-threatening toxicity given that patient had a complete response. Therefore, (p_{Ro}, p_{To}) is assumed to be $(0.50, 0.30)$, and the odds ratio is assumed to be 2.0. Conaway and Petroni¹ chose values $\alpha = 0.05$, $\gamma = 0.30$, and $\beta = 0.10$. The trial is designed to have approximately 90% power at the alternative determined by $(p_{Ra}, p_{Ta}) = (0.75, 0.15)$.

The extension to multi-stage designs is straightforward. The multi-stage designs allow for the early termination of a study if early results indicate that the treatment is not sufficiently effective or is too toxic. Although most phase II trials are carried out in at most two stages, for the general discussion, Conaway and Petroni¹ assume that the study is to be carried out in K stages. At the end of the k th-stage, a decision is made whether to enroll patients for the next stage or to stop the trial. If the trial is stopped early, the treatment is declared not sufficiently promising to warrant further study. At the end of the k th stage, the decision to continue or terminate the study is governed by the boundaries (c_{Rk}, c_{Tk}) , $k = 1, \dots, K$. The study continues to the next stage if the total number of responses observed up to and including the k th stage is at least as great as c_{Rk} and the total number of toxicities up to and including the k th stage is no greater than c_{Tk} . At the final stage, the null hypothesis that the treatment is not sufficiently promising to warrant further study is rejected if there are a sufficient number of observed responses (at least c_{RK}) and sufficiently few observed toxicities (no more than c_{TK}).

In designing the study, the goal is to choose sample sizes for the stages $m_1, m_2, \dots, m_K$ and boundaries $(c_{R1}, c_{T1}), (c_{R2}, c_{T2}), \dots, (c_{RK}, c_{TK})$ satisfying the error constraints listed above. For a fixed total sample size, $N = \sum_k m_k$, there may be many designs that satisfy the error requirements. An additional criterion, such as one of those proposed by Simon³ in the context of two-stage trials with a single binary endpoint, can be used to select a design. The stage sample sizes and boundaries can be chosen to give the minimum expected sample size at the response and toxicity rates for the standard therapy (p_{Ro}, p_{To}) among all designs that satisfy the error requirements. Alternatively, one could choose the design that minimizes the maximum expected sample size over the entire null hypothesis region.

Conaway and Petroni¹ compute the optimal designs for these criteria for two-stage and three-stage designs using a fixed prespecified value for the odds ratio θ . Through simulations, they evaluate the sensitivity of the designs to a misspecification of the value for the odds ratio.

Bryant and Day² also consider the problem of monitoring binary endpoints representing response and toxicity. They present optimal designs for two-stage trials that extend the designs of Simon.³ In the first stage, N_1 patients are accrued and classified by response and toxicity; Y_{R1} patients respond, and Y_{T1} patients do not experience toxicity. At the end of the first stage, a decision to continue to the next stage or to terminate the study is made according to the following rules, where N_1 , C_{R1} , and C_{T1} are parameters to be chosen as part of the design specification:

1. If $Y_{R1} \leq C_{R1}$ and $Y_{T1} > C_{T1}$, terminate due to inadequate response.
2. If $Y_{R1} > C_{R1}$ and $Y_{T1} \leq C_{T1}$, terminate due to excessive toxicity.
3. If $Y_{R1} \leq C_{R1}$ and $Y_{T1} \leq C_{T1}$, terminate due to both factors.
4. If $Y_{R1} > C_{R1}$ and $Y_{T1} > C_{T1}$, continue to the second stage.

In the second stage, $N_2 - N_1$ patients are accrued. At the end of this stage, the following rules govern the decision whether or not the new agent is sufficiently promising, where N_2 , C_{R2} , and C_{T2} are parameters to be determined by the design and Y_{R2} patients respond and Y_{T2} patients do not experience toxicity over all first and second stage patients combined:

1. If $Y_{R2} \leq C_{R2}$ and $Y_{T2} > C_{T2}$, not promising due to inadequate response.
2. If $Y_{R2} > C_{R2}$ and $Y_{T2} \leq C_{T2}$, not promising due to excessive toxicity.
3. If $Y_{R2} \leq C_{R2}$ and $Y_{T2} \leq C_{T2}$, not promising due to both factors.
4. If $Y_{R2} > C_{R2}$ and $Y_{T2} > C_{T2}$, sufficiently promising.

The principle for choosing the stage sample sizes and stage boundaries is the same as in Conaway and Petroni;¹ the design parameters are determined from prespecified error constraints. Although Bryant and Day² and Conaway and Petroni¹ differ in the particular constraints considered, the motivation for these error constraints is the same. One would like to limit the probability of recommending a treatment that has an insufficient response rate or excessive toxicity rate. Similarly, one would like to constrain the probability of failing to recommend a treatment that is superior to the standard treatment in terms of both response and toxicity rates. Finally, among all designs meeting the error criteria, the optimal design is the one that minimizes the average number of patients treated with an ineffective therapy.

In choosing the design parameters, $Q = (N_1, N_2, C_{R1}, C_{R2}, C_{T1}, C_{T2})$, Bryant and Day² specify an acceptable (P_{R1}) and an unacceptable (P_{R0}) response rate along with an acceptable (P_{T1}) and unacceptable (P_{T0}) rate of non-toxicity. Under any of the four combinations of acceptable or unacceptable rates of response and non-toxicity, Bryant and Day² assume that the association between response and toxicity is constant. The association between response and toxicity is determined by the odds ratio ϕ in the 2×2 table cross-classifying response and toxicity,

$$\phi = \frac{P(\text{No Response, Toxicity}) * P(\text{Response, No Toxicity})}{P(\text{No Response, No Toxicity}) * P(\text{Response, Toxicity})}$$

Bryant and Day² parameterize the odds ratio in terms of response and no toxicity so ϕ corresponds to $1/\theta$ in the notation of Conaway and Petroni.¹ For a design Q and an odds ratio ϕ , let $\alpha_{ij}(Q, \phi)$ be the probability of recommending the treatment, given that the true response rate equals P_{Ri} and the true non-toxicity rate equals P_{Tj} , $i = 0, 1$; $j = 0, 1$. Constraining the probability of recommending a treatment with an insufficient response rate leads to $\alpha_{01}(Q, \phi) \leq \alpha_R$, where α_R is a pre-specified constant. Constraining the probability of recommending an overly toxic treatment leads to $\alpha_{10}(Q, \phi) \leq \alpha_T$, and ensuring a sufficiently high probability of recommending a truly superior treatment requires $\alpha_{11}(Q, \phi) \geq 1 - \beta$, where α_T and β are pre-specified constants. Bryant and Day² note that $\alpha_{00}(Q, \phi)$ is less than either $\alpha_{01}(Q, \phi)$ or $\alpha_{10}(Q, \phi)$, so that an upper bound on $\alpha_{00}(Q, \phi)$ is implicit in these constraints.

There can be many designs that meet these specifications. Among these designs, Bryant and Day² define the optimal design to be the one that minimizes the expected number of patients in a study of a treatment with an unacceptable response or

toxicity rate. Specifically, Bryant and Day² choose the design Q that minimizes the maximum of $E_{01}(Q, \phi)$ and $E_{10}(Q, \phi)$, where E_{ij} is the expected number of patients accrued when the true response rate equals P_{Ri} and the true nontoxicity rate equals P_{Tj} , $i = 0, 1$; $j = 0, 1$. The expected value $E_{00}(Q, \phi)$ does not play a role in the calculation of the optimal design because it is less than both $E_{01}(Q, \phi)$ and $E_{10}(Q, \phi)$.

The stage sample sizes and boundaries for the optimal design depend on the value of the nuisance parameter ϕ . For an unspecified odds ratio, among all designs that meet the error constraints, the optimal design minimizes the maximum expected patient accruals under a treatment with an unacceptable response or toxicity rate, $\max_{\phi} \{ \max(E_{01}(Q, \phi), E_{10}(Q, \phi)) \}$. Assumptions about a fixed value of the odds ratio lead to a simpler computational problem; this is particularly true if response and toxicity are assumed to be independent ($\phi = 1$). Bryant and Day² provide bounds that indicate that the characteristics of the optimal design for an unspecified odds ratio do not differ greatly from the optimal design found by assuming that response and toxicity are independent. By considering a number of examples, Conaway and Petroni¹ came to a similar conclusion. Their designs are computed under a fixed value for the odds ratio, but different values for the assumed odds ratio led to similar designs.

7.3 DESIGNS THAT ALLOW A TRADE-OFF BETWEEN RESPONSE AND TOXICITY

The designs for response and toxicity proposed by Conaway and Petroni¹ and Bryant and Day² share a number of common features, including the form for the alternative region. In these designs, a new treatment must show evidence of a greater response rate and a lesser toxicity rate than the standard treatment. In practice, a trade-off could be considered in the design, since one may be willing to allow greater toxicity to achieve a greater response rate or may be willing to accept a slightly lower response rate if lower toxicity can be obtained. Conaway and Petroni⁴ propose two-stage designs for phase II trials that allow for early termination of the study if the new therapy is not sufficiently promising and allow for a trade-off between response and toxicity.

The hypotheses are the same as those considered for the bivariate designs of the previous section. The null hypothesis is that the new treatment is not sufficiently promising to warrant further study, due either to an insufficient response rate or excessive toxicity. The alternative hypothesis is that the new treatment is sufficiently effective and safe to warrant further study. The terms *sufficiently safe* and *sufficiently effective* are relative to the response rate p_{Ro} and the toxicity rate p_{To} for the standard treatment.

One of the primary issues in the design is how to elicit the trade-off specification. Ideally, the trade-off between safety and efficacy would be summarized as a function of toxicity and response rates that defines a treatment as worthy of further study. In practice this can be difficult to elicit. A simpler method for obtaining the trade-off information is for the investigator to specify the maximum toxicity rate $p_{T,max}$ that would be acceptable if the new treatment were to produce responses in all patients. Similarly, the investigator would be asked to specify the minimum response rate $p_{R,min}$ that would be acceptable if the treatment produced no toxicities.

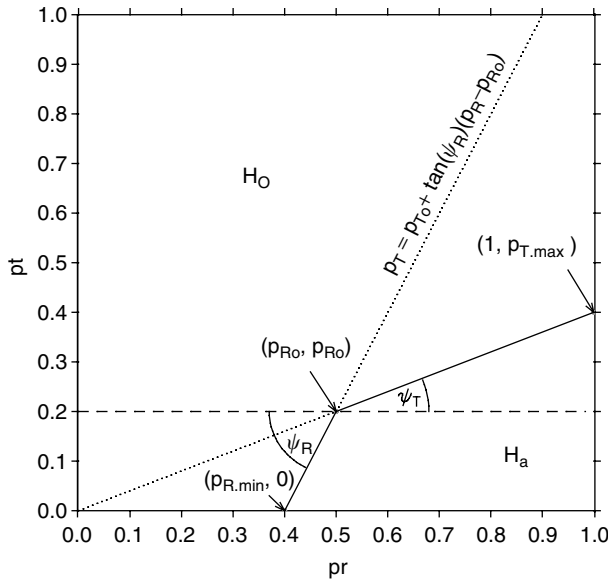


FIGURE 7.2 Null and alternative regions for trade-offs.

Figure 7.2 illustrates the set of values for the true response rate (p_R) and true toxicity rate (p_T) that satisfy the null and alternative hypotheses. The values chosen for Figure 7.2 are $p_{R0} = 0.5$, $p_{T0} = 0.2$, $p_{R,min} = 0.4$, and $p_{T,max} = 0.7$. The line connecting the points (p_{R0}, p_{T0}) and $(1, p_{T,max})$ is given by the equation $p_T = p_{T0} + \tan(\psi_T)(p_R - p_{R0})$, where $\tan(\psi_T) = (p_{T,max} - p_{T0})/(1 - p_{R0})$. Similarly, the equation of the line connecting (p_{R0}, p_{T0}) and $(1, p_{R,min})$ is given by the equation $p_T = p_{T0} + \tan(\psi_R)(p_R - p_{R0})$, where $\tan(\psi_R) = p_{T0}/(p_{R0} - p_{R,min})$. With $\psi_T \leq \psi_R$ the null hypothesis is

$$H_0: p_T \geq p_{T0} + \tan(\psi_T)(p_R - p_{R0}) \text{ or } p_T \geq p_{T0} + \tan(\psi_R)(p_R - p_{R0})$$

and the alternative hypothesis is

$$H_a: p_T < p_{T0} + \tan(\psi_T)(p_R - p_{R0}) \text{ and } p_T < p_{T0} + \tan(\psi_R)(p_R - p_{R0})$$

The forms of the null and alternative are different for the case where $\psi_T \geq \psi_R$, although the basic principles in constructing the design and specifying the trade-off information remain the same (cf., Conaway and Petroni).⁴ Special cases of these hypotheses have been used previously: $\psi_T = 0$ and $\psi_R = \pi/2$ yield the critical regions of Conaway and Petroni¹ and Bryant and Day;² $\psi_R = \psi_T = 0$ yield hypotheses in terms of toxicity alone; and $\psi_R = \psi_T = \pi/2$ yield hypotheses in terms of response alone.

To describe the trade-off designs for a fixed sample size, we use the notation and assumptions for the fixed sample size design described in Section 7.2. As in their earlier work, Conaway and Petroni⁴ determine sample size and critical values under an assumed value for the odds ratio between response and toxicity. The sample size

calculations require a specification of a level of type I error α and power $1-\beta$, at a particular point $p_R = p_{Ra}$ and $p_T = p_{Ta}$. The point (p_{Ra}, p_{Ta}) satisfies the constraints defining the alternative hypothesis and represents the response and toxicity rates for a treatment considered to be superior to the standard treatment. The test statistic is denoted by $T(\hat{p})$, where $\hat{p} = (1/N)(X_{11}, X_{12}, X_{21}, X_{22})$ is the vector of sample proportions in the four cells of Table 7.1, and is based on computing an I-divergence measure (cf., Robertson, Dykstra, and Wright).⁸

The test statistic has the intuitively appealing property of being roughly analogous to a distance from $\hat{p}$ to the region H_o . Rejection of the null hypothesis results when the observed value of $T(\hat{p})$ is too far from the null hypothesis region. A vector of observed proportions $\hat{p}$ leads to rejection of the null hypothesis if $T(\hat{p}) \geq c$. For an appropriate choice of sample size (N), significance level (α), and power ($1-\beta$), the value c can be chosen to: 1) constrain the probability of recommending a treatment that has an insufficient response rate relative to the toxicity rate, and 2) ensure that there is a high probability of recommending a treatment with response rate p_{Ra} and toxicity rate p_{Ta} . The critical value c is chosen to meet the error criteria:

$$\sup_{(P_R, P_T) \in H_o} P(T(\hat{p}) \geq c \mid p_R, p_T, \theta) \leq \alpha \quad (7.4)$$

and

$$P(T(\hat{p}) \geq c \mid p_{Ra}, p_{Ta}, \theta) \geq 1 - \beta \quad (7.5)$$

These probabilities are computed for a fixed value of the odds ratio θ by enumerating the value of $T(\hat{p})$ for all possible realizations of the multinomial vector $(X_{11}, X_{12}, X_{21}, X_{22})$.

The trade-off designs can be extended to two-stage designs that allow for early termination of the study if the new treatment does not appear to be sufficiently promising. In designing the study, the goal is to choose the stage sample sizes (m_1, m_2) and decision boundaries (c_1, c_2) to satisfy error probability constraints similar to those in the fixed sample size trade-off design:

$$\sup_{(P_R, P_T) \in H_o} P(T_1(\hat{p}_1) \geq c_1, T_2(\hat{p}_1, \hat{p}_2) \geq c_2 \mid p_R, p_T, \theta) \leq \alpha \quad (7.6)$$

and

$$P(T_1(\hat{p}_1) \geq c_1, T_2(\hat{p}_1, \hat{p}_2) \geq c_2 \mid p_{Ra}, p_{Ta}, \theta) \geq 1 - \beta \quad (7.7)$$

where T_1 is the test statistic computed on the stage 1 observations and T_2 is the test statistic computed on the accumulated data in stages 1 and 2. As in the fixed sample size design, these probabilities are computed for a fixed value of the odds ratio and are found by enumerating all possible outcomes of the trial.

In cases where many designs meet the error requirements, an optimal design is found according to the criterion in Bryant and Day² and Simon.³ Among all designs that meet the error constraints, the chosen design minimizes the maximum expected sample size under the null hypothesis. Through simulations, Conaway and Petroni⁴

investigate the effect of fixing the odds ratio on the choice of the optimal design. They conclude that unless the odds ratio is badly misspecified, the choice of the odds ratio has little effect on the properties of the optimal design.

The critical values for the test statistic are much harder to interpret than the critical values in Conaway and Petroni¹ or Bryant and Day,² which are counts of the number of observed responses and toxicities. We recommend two plots similar to Figures 2 and 3 in Conaway and Petroni⁴ to illustrate the characteristics of the trade-off designs. The first is a display of the power of the test, so that the investigators can see the probability of recommending a treatment with true response rate p_R and true toxicity rate p_T . The second plot displays the rejection region, so that the investigators can see the decision about the treatment that will be made for specific numbers of observed responses and toxicities. With these plots, the investigators can better understand the implications of the trade-off being proposed.

The trade-off designs of Conaway and Petroni⁴ were motivated by the idea that a new treatment could be considered acceptable even if the toxicity rate for the new treatment is greater than that of the standard treatment, provided the response rate improvement is sufficiently large. This idea also motivated the Bayesian monitoring method of Thall, Simon, and Estey.^{5,6} They note that, for example, a treatment that improves the response rate by 15 percentage points might be considered promising, even if its toxicity rate is 5 percentage points greater than the standard therapy. If, however, the new therapy increases the toxicity rate by 10 percentage points, it might not be considered an acceptable therapy.

Thall, Simon, and Estey^{5,6} outline a strategy for monitoring each endpoint in the trial. They define for each endpoint in the trial a monitoring boundary based on pre-specified targets for an improvement in efficacy and an unacceptable increase in the rate of adverse events. In the example given above for a trial with a single response endpoint and a single toxicity endpoint, the targeted improvement in response rate is 15%, and the allowance for increased toxicity is 5%.

Thall, Simon, and Estey^{5,6} take a Bayesian approach that allows for monitoring each endpoint on a patient by patient basis. Although their methods allow for a number of efficacy and adverse event endpoints, we will simplify the discussion by considering only a single efficacy event (response) and a single adverse event endpoint (toxicity). Before the trial begins, they elicit a prior distribution on the cell probabilities in Table 7.2. Under the standard therapy, the cell probabilities are denoted $P_S = (p_{S11}, p_{S12}, p_{S21}, p_{S22})$; under the new experimental therapy, the cell probabilities are denoted $P_E = (p_{E11}, p_{E12}, p_{E21}, p_{E22})$. Putting a prior distribution on the cell probabilities $(p_{G11}, p_{G12}, p_{G21}, p_{G22})$ induces a prior distribution on $p_{GR} = p_{G11} + p_{G12}$ and on $p_{GT} = p_{G11} + p_{G21}$, where G stands for either S or E . A Dirichlet prior for the cell probabilities is particularly convenient in this setting, since this induces a beta prior on p_{GR} and p_{GT} for $G = S$ or E .

In addition to the prior distribution, Thall, Simon, and Estey^{5,6} specify a target improvement $\delta(R)$ for response and a maximum allowable difference $\delta(T)$ for toxicity. The monitoring of the endpoints begins after a minimum number of patients m have been observed. It continues until either a maximum number of patients M have been accrued or a monitoring boundary has been crossed.

In a typical phase II trial, in which only the new therapy is used, the distribution on the probabilities under the standard therapy remains constant throughout the trial,

while the distribution on the probabilities under the new therapy is updated each time a patient's outcomes are observed. After the response and toxicity classification on j patients, X_j , have been observed, there are several possible decisions one could make. If there is strong evidence that the new therapy does not meet the targeted improvement in response rate, then the trial should be stopped and the new treatment declared not sufficiently promising. Alternatively, if there is strong evidence that the new treatment is superior to the standard treatment in terms of the targeted improvement for response, the trial should be stopped and the treatment declared sufficiently promising. In terms of toxicity, the trial should be stopped if there is strong evidence of an excessive toxicity rate with the new treatment. Thall, Simon, and Estey^{5,6} translate these rules into statements about the updated (posterior) distribution $[p_E | X_j]$ and the prior distribution p_S using pre-specified cutoffs for what constitutes strong evidence. For $m \leq j \leq M$, the monitoring boundaries are:

$$P[p_{ER} - p_{SR} > \delta(R) | X_j] \leq p_L(R) \quad (7.8)$$

$$P[p_{ER} > p_{SR} | X_j] \geq p_U(R) \quad (7.9)$$

$$P[p_{ET} - p_{ST} > \delta(T) | X_j] \geq p_U(T) \quad (7.10)$$

where $p_L(R)$, $p_U(R)$, and $p_U(T)$ are prespecified probability levels. Numerical integration is required to compute these probabilities, but the choice of the Dirichlet prior makes the computations relatively easy. Extensions to the method that allow for mixture priors and monitoring cohorts of size greater than one are given in Thall and Sung.⁷

Thall and Russell⁹ present Bayesian methods for combined phase I/II trials. These designs can be used for dose finding based on response and toxicity criteria. The models impose an ordering on a combined response-toxicity endpoint and monitor the trial by updating the probability of response and toxicity.

7.4 SUMMARY

All of the methods discussed in this chapter have advantages in monitoring toxicity in phase II trials. None of the methods use asymptotic approximations for distributions, and all are well suited for the small sample sizes encountered typically in phase II trials. The bivariate designs of Conaway and Petroni¹ and Bryant and Day² have critical values that are based on the observed number of responses and the observed number of toxicities; these statistics are easily calculated and interpreted by the investigators. Although there is no formal trade-off discussion in these papers, the general methods can be adapted to the kind of trade-off discussed in Simon, Thall, and Estey.^{5,6} To do this, one needs to modify the hypotheses to be tested. For example, the null and alternative hypothesis could be changed to

$$H_o: p_R \leq p_{Ro} + \delta_R \text{ or } p_T \geq p_{To} + \delta_T$$

$$H_a: p_R > p_{Ro} + \delta_R \text{ and } p_T < p_{To} + \delta_T$$

for some prespecified δ_R and δ_T . The trade-off designs of Conaway and Petroni⁴ have a trade-off strategy that permits the allowable level of toxicity to increase with the response rate. In contrast, in the trade-off example of Thall, Simon, and Estey,^{5,6} a 5% increase in toxicity would be considered acceptable for a treatment with a 15% increase in response. Because the allowance in toxicity is prespecified, this means that only a 5% increase in toxicity is allowable even if the response rate with the new treatment is as much as 30%. With the trade-off of Conaway and Petroni,⁴ the standard for allowable toxicity is greater for a treatment with a 30% improvement than for one with a 15% improvement. The methods of Thall, Simon, and Estey^{5,6} have advantages in terms of being able to monitor outcomes on a patient by patient basis. At each monitoring point, the method can provide graphical representations of the probability associated with each of the decision rules.

Although the methods presented are discussed in terms of toxicity and response, where toxicity is a predefined measure of adverse events related to protocol treatment and response is a predefined measure of efficacy, the designs apply to any bivariate endpoints. For example, in vaccine trials assessing immune response, the efficacy response parameter could be replaced with an earlier measure of immune response.

REFERENCES

1. Conaway MR and Petroni GR. Bivariate sequential designs for phase II trials. *Biometrics*. 1995;51:656–664.
2. Bryant J and Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics*. 1995;51:1372–1383.
3. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*. 1989;10:1–10.
4. Conaway MR and Petroni GR. Designs for phase II trials allowing for trade-off between response and toxicity. *Biometrics*. 1996;52:1375–1386.
5. Thall PF, Simon RM, and Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in Medicine*. 1995;14:357–379.
6. Thall PF, Simon RM, and Estey EH. New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. *Journal of Clinical Oncology*. 1996;14:296–303.
7. Thall PF and Sung HG. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Statistics in Medicine*. 1998;17:1563–1580.
8. Robertson T, Dykstra RL, and Wright FT. *Order Restricted Statistical Inference*. Chichester: John Wiley and Sons, Ltd., 1988.
9. Thall PF and Russell KE. A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. *Biometrics*. 1998;54:251–264.

8 Phase II Trials Using Time-to-Event Endpoints

*Catherine M. Tangen and
John J. Crowley*

CONTENTS

8.1	Introduction	143
8.2	Response Is Not Strongly Correlated with Survival	144
8.3	Challenges of Evaluating Best Response in Cancer Clinical Trials	146
8.3.1	Requiring Measurable Disease	146
8.3.2	Difficulty Assessing Response	146
8.3.3	Cytostatic Agents	146
8.4	Phase II Designs Using Time to Event Endpoints	146
8.4.1	Single Arm Phase II Designs	147
8.4.1.1	Comparisons with Historical Experience	147
8.4.1.2	Each Patient as His Own Control	148
8.4.2	Multi-Arm Designs.....	150
8.4.2.1	Randomized Selection Designs	150
8.4.2.2	Comparison of Agent with Placebo.....	150
8.4.2.3	Randomization Discontinuation Design	150
8.5	Concluding Remarks	153
	References	153

8.1 INTRODUCTION

The objective of a phase II study is to evaluate whether a particular regimen has enough biologic activity in a given disease to warrant further investigation. We would like to have a mechanism where the candidate agents or regimens can be screened relatively quickly, and for practical and ethical reasons we would like to expose the minimal number of patients in order to evaluate activity.

Typically, phase II cancer clinical trials use response rate as the primary endpoint, where in the solid tumor setting, response usually involves a reduction in the dimensions of the measurable disease. However, there are shortcomings to using response rate as the primary endpoint for evaluating efficacy: (1) Response has not been shown to be strongly correlated with survival across most disease sites, (2) there are challenges to

evaluating response, and (3) there are new classes of agents that are being tested that are not expected to be cytotoxic (tumor reducing) but instead are cytostatic; that is, the agents may delay disease progression but not reduce the tumor size. For these reasons, other endpoints such as survival or progression-free survival should be considered as the primary endpoint for evaluating phase II drugs.

This chapter will briefly address some of the shortcomings and challenges involved with using response as a primary endpoint for phase II trials. We will then describe a number of proposed phase II designs that use time-to-event endpoints both in the setting of single-arm and multiple-arm trials. The merits and drawbacks of these proposed alternative phase II designs will be discussed.

8.2 RESPONSE IS NOT STRONGLY CORRELATED WITH SURVIVAL

A therapy may result in objective responses, but it may not improve survival. There have been a number of examples of this in the clinical literature. For example, this was seen in a phase III advanced squamous-cell head and neck study that was conducted by the Southwest Oncology Group. This three-arm randomized phase III study compared cisplatin plus fluorouracil (CDDP + 5-FU in Figure 8.1) and carboplatin plus fluorouracil (CBDCA + 5-FU in Figure 8.1) to single-agent methotrexate.¹ The response rates were 32%, 21%, and 10%, respectively. The response rate comparison of cisplatin plus 5-FU to methotrexate was statistically significant ($P<0.001$). However, as can be seen in Figure 8.1, the survival curves are virtually identical for the three regimens.

There are two reasonable analytic strategies for testing the hypothesis that responders have better survival than nonresponders. The first involves using a time-dependent covariate for response to indicate when a patient moves from a non-responder state.² The other approach is to perform a landmark analysis for all patients at some fixed time following the onset of treatment. For a phase II study Anderson

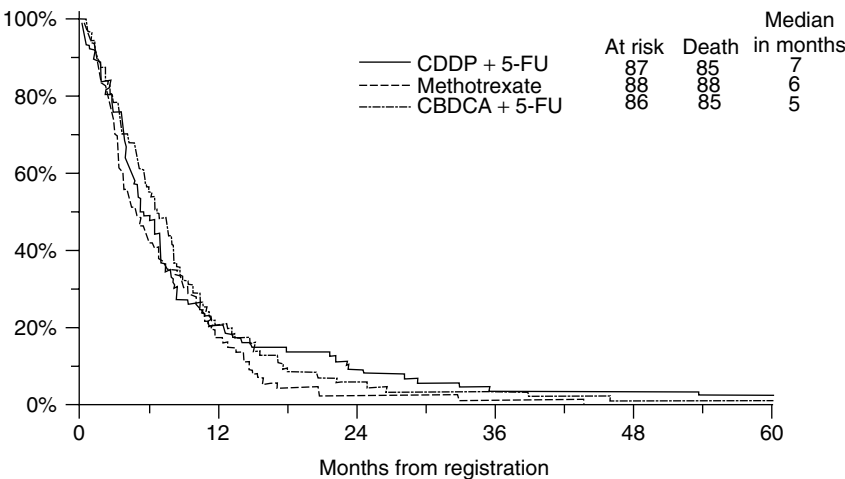


FIGURE 8.1 Survival curves by the treatment arm for a randomised phase III trial in head and neck cancer.

et al.³ suggest the landmark might be at a time at which most patients who are going to respond have already done so. Patients who progress or who are “off-study” prior to the landmark are excluded from the analysis. Patients are analyzed according to their response status at the time of the landmark.

Durie et al.⁴ examined four phase III standard chemotherapy multiple myeloma trials with 1555 patients to clarify the predictive value of specific levels of M-component (monoclonal protein) or M-component reduction ($\geq 50\%$ and $\geq 75\%$) as a response definition vs. time to progression in assessment of treatment benefit. Progression was defined as a 25% increase over baseline in the calculated tumor mass or other signs of disease progression such as hypercalcemia. In responding patients, progression was defined as the first occurrence of any of the following: an increase in tumor mass greater than 100% over the lowest level recorded, an increase to greater than the levels defining response, increase of lytic bone lesions, or evidence of new disease activity. Using six- and twelve-month landmarks, they found that survival was not different for responders versus nonresponders in patients without disease progression prior to the landmark. Conversely, progression of disease had a clear impact on survival. A proportional hazards regression model was fit to overall survival, with response and disease progression placed in the model as time-dependent covariates. As can be seen in Table 8.1, survival duration was influenced more by progression than by response. However, the authors did note that patients who progressed after a response were at greater risk of death than those who had not had a prior response, as can be seen by the significant interaction term. The authors’ recommendation was that new therapies should be evaluated in terms of time to first progression rather than precise M-component reduction levels.

TABLE 8.1

Univariate and Multivariate Proportional Hazards Regression Models for Response and/or Progression Predicting Subsequent Survival in Patients with Multiple Myeloma. HR indicates Cox model hazard ratio; p, Cox model p-value; all models stratified by protocol

	Six-Month Landmark (n = 1376)			One-Year Landmark Landmark (n = 1231)			Time-Dependent Model (n = 1555)		
	%	HR	P	%	HR	P	%	HR	P
Univariate									
Progression	6%	2.09	<0.001	18%	1.74	<0.001	55%	3.22	<0.001
$\geq 75\%$ remission	31%	1.01	0.790	41%	0.83	0.002	43%	0.79	<0.001
Any response	46%	1.12	0.067	59%	0.80	<0.001	62%	0.81	0.002
Multivariate									
Progression	6%	1.46	0.018	18%	1.56	<0.001	55%	2.65	<0.001
$\geq 75\%$ remission	31%	1.04	0.579	41%	0.84	0.040	43%	0.66	<0.001
50–75% remission	15%	1.16	0.112	18%	0.94	0.532	19%	0.83	0.007
Interaction*	1.5%	1.26	<0.001	10%	1.33	<0.001	40%	1.39	<0.001

* *Interaction* term is progression $\times$ any response.

Source: Durie BGM, Jacobson J, Barlogie B, Crowley J, *J. Clin. Onc.* 22:1857–1863, 2004, with permission from the American Society of Clinical Oncology.

8.3 CHALLENGES OF EVALUATING BEST RESPONSE IN CANCER CLINICAL TRIALS

8.3.1 REQUIRING MEASURABLE DISEASE

In order to evaluate response, patients must have measurable disease. In some diseases, measurable objective disease is often not present. For example, most advanced prostate cancer patients present with bone lesions (nonmeasurable disease), and measurable disease is relatively rare. In order to evaluate objective response as the primary endpoint, a study must restrict participation to those patients with measurable disease, and it may be that this subset of individuals is not representative of the disease population. In the prostate cancer setting, investigators have tried to address this by defining measurable disease as a minimum level of PSA at study entry.

8.3.2 DIFFICULTY ASSESSING RESPONSE

Even if a patient does have measurable disease, it is often challenging to get complete and timely measurements on all disease during follow-up assessments. This can be due to rapid deterioration of the patient's condition, financial considerations for the cost of frequent scans, or difficulty reading and measuring the images.⁵⁻⁸ When patients are not assessable for response due to inadequate disease information, the conservative approach is to consider these individuals to be nonresponders. However, in some cases this may be inappropriate. In sarcoma, for example, an individual might have numerous disease sites at study entry, and it is not uncommon for many of those sites to disappear with treatment. However, if the original sites are not evaluated post-randomization and it is explicitly noted that the disease has disappeared, these patients will be considered not assessable for response. In that case, it would be misleading to categorize inadequate disease assessment as a nonresponse.

8.3.3 CYTOSTATIC AGENTS

There are new classes of target agents that may lengthen the progression-free interval or survival but yet not produce a tumor response. Antiangiogenesis factors, growth modulators, monoclonal antibodies, and gene therapy vaccines are all being tested in phase II studies. Instead of killing cells, they change the environment or they attack specific cell targets. For example, angiogenesis factors block the growth of blood supply, thereby not allowing the tumor to grow.⁹ These mechanisms are not necessarily expected to shrink tumors but instead are more likely to delay progression. Using tumor response as the primary endpoint for these types of drugs may fail to identify truly active agents.¹⁰ There is also not a clear distinction between agents that have cytostatic versus cytotoxic mechanisms of action. Alternatives to tumor response should be considered for a wide range of drugs.

8.4 PHASE II DESIGNS USING TIME TO EVENT ENDPOINTS

Using survival or progression-free survival as the primary endpoint of a phase II study has appeal because it measures (or more closely measures) a clinically meaningful

endpoint that impacts a patient's life. It also provides consistency with planned phase III trials that would also use survival or progression-free survival as the primary endpoint. Two arguments that have been made in support of using response as an endpoint are that (1) subsequent treatments do not impact it, and (2) response status is evaluated more quickly than progression or survival status. However, by using progression-free survival or survival as the primary endpoint, it is not necessary to wait until all patients have progressed or died or even until the median has been estimated. It is possible to specify the null and alternative hypotheses in terms of a shorter interval of time, such as the expected event rate at six months, which should not be impacted much by subsequent treatments.

8.4.1 SINGLE ARM PHASE II DESIGNS

8.4.1.1 Comparisons with Historical Experience

This strategy is typically used for phase II studies that have response as the primary endpoint. Based on previous experience, one would specify the level at which there would not be interest in an agent (null hypothesis) versus the alternative hypothesis, which is the level at which one would consider pursuing an agent in a phase III study. For example, one might specify a 20% vs. 40% response rate for an agent where there was some modest level of activity using standard treatment. A two-stage design is then typically employed, where if adequate activity is observed in the first group of patients, the study will continue to the second stage of accrual;^{11,12} see also the chapter by S. Green in this volume. If the regimen being studied consists of agents already shown to be active, a single-stage (or pilot) design may be appropriate. Because there is already experience with the agents being studied, there is less concern about exposing patients to an ineffective treatment, so a one-stage design is usually used with a sample size of typically 50 to 100 patients, avoiding the delays of a multistage design.

A similar design strategy could be used in the phase II setting by substituting progression-free survival or survival for response rate. One could specify a 6 month progression-free survival estimate of 20% versus 40%, for example, where if only 20% of the patients are alive and free of progression at 6 months, there would be no interest in the regimen, whereas if 40% or more of the patients are free of progression, then there would be considerable interest in pursuing the agent if other considerations such as toxicity were also favorable. This could be tested in a one- or two-stage design.

One potential criticism for using progression-free or overall survival as a primary endpoint may be that patients have to be followed for a longer period of time to obtain the result, especially if the estimate is based on a 1 year interval instead of 6 months. However, particularly in cooperative groups or other multi-institutional settings, it is not unusual for full response assessment also to take a long time to evaluate. In fact, in the advanced disease setting, progression is often assessed prior to obtaining complete information on the best response for an individual patient. Nevertheless, it is likely that for a two-stage design, it may be necessary to temporarily close the study to accrual while endpoint status is being assessed for the first stage.

One design to avoid temporary closure was proposed by Herndon,¹³ where a slight over-accrual to the first stage is allowed while assessing the endpoint on

patients in the first cohort. In a similar vein, it may be reasonable to conduct an interim analysis of a phase II study in order to avoid temporary closure. The alternative hypothesis is tested at the alpha 0.005 level or at some other appropriate level at the approximate midpoint, and the study would be closed only for the case where there is lack of adequate activity.

One straightforward design option for conducting a pilot study would be to specify the null and alternative hypotheses in terms of median progression-free or overall survival. For example, in the advanced disease setting a median survival of 9 months might not be of interest, while a median survival of 12 months or greater would be of further interest. By assuming uniform accrual and an exponential distribution, it is straightforward to calculate the sample size needed. If a study is to have 90% power and take two years of accrual with one additional year of follow-up and a one-sided $\alpha = 0.05$, then 134 patients, would be needed. With this one-stage design it also is possible to specify an interim analysis to test for lack of biologic activity.

One of the challenges of a design that makes comparisons with historical experience is choosing the appropriate null and alternative hypothesis levels. As recommended by Korn et al.,¹⁴ there must be sufficient historical data on a patient population, untreated or treated with active agents that are similar to the patient population being considered for treatment with the experimental agent. The historical data would need to be the survival or progression-free survival experience for a group of patients with the same stage of disease and amount of prior treatment and similar organ function and performance status, and the procedures used for monitoring progression should be the same. Another important recommendation is that patients should come from the same type of institutions with the same referral patterns in a recent era so diagnostic measures and supportive care would be similar. For example, using the results of a single institution study in order to define the level of interest and disinterest in a regimen might not be readily translatable to a large, diverse cooperative group phase II trial. It should be pointed out that these are essentially the same factors that need to be considered when designing a phase II *response* study using historical experience as a comparison.

8.4.1.2 Each Patient as His Own Control

With this type of design in a single group of patients who have progressive disease, we want to evaluate whether an agent is able to slow the rate of progression relative to the patients' pretreatment rate of progression.

Mick et al.¹⁵ have proposed a methodology for evaluating time-to-progression as the primary endpoint in a one-stage design. Typically, patients being offered phase II studies of new agents have failed a previous regimen. The prior time to progression interval is referred to as TTP_1 and is not censored; that is, all progressions are observed. Time to progression after the cytostatic agent, TTP_2 , may or may not be censored at analysis. They propose that the growth modulation index (TTP_2/TTP_1) will have a null ratio value (HR_0) of 1.0, and the index needs to be greater than 1.33 if a cytostatic agent is to be considered effective at delaying progression. The degree of correlation between the paired failure times is a key feature of this design because the patient serves as his

own historical control, a concept that was originally suggested by Von Hoff.¹⁶ The authors note that in some cases it may be reasonable to hypothesize that by the natural history of the disease one would expect TTP_2 to be shorter than TTP_1 , which would indicate a null value less than 1.0. Hypothesis tests about the hazard ratio from paired data may be conducted under a log-linear model.¹⁷ A test of $HR_0 = 1$ is equivalent to a test of $\ln(HR_0)$ equal to zero. Using a log-linear model where β_0 denotes $\log(HR_0)$, the values of natural log-transformed paired failure times are compared. Failure-time pairs in which $\ln(TTP_2) \geq \ln(TTP_1) + \beta_0$ and for which $\ln(TTP_2) < \ln(TTP_1) + \beta_0$ and for which TTP_2 is uncensored are scored as 1 or -1, respectively. Pairs in which $\ln(TTP_2) < \ln(TTP_1) + \beta_0$ and for which TTP_2 is censored do not contribute to the test. The test statistic is a function of sums of squares of the scores for contributing failure-time pairs and has an approximate chi-square distribution. Table 8.2 provides examples of statistical power that is attained for a range of sample sizes, alternative hypotheses, and correlations of the paired failure time data.

One can propose a range of null and alternative hypotheses based on disease and treatment considerations. However, this design approach can be difficult to implement because patients must be enrolled for both intervals of progression; that is, patients are enrolled prior to their first-line treatment for a trial of second-line treatment. As pointed out by Korn et al.,¹⁴ enrolling patients after they progress on first-line treatment avoids these problems but leads to potential bias in the selection of the patients included in the trial.

TABLE 8.2
Power Based on Effect Size^a, Study Size^b, and Correlation^c

$H_0: \Delta_0 = 1$ $H_A: \Delta_A =$	$\rho =$	30 Patients	40 Patients	50 Patients	60 Patients	70 Patients
1.1	0.3	0.053	0.079	0.083	0.078	0.107
	0.5	0.112	0.120	0.123	0.147	0.152
	0.7	0.239	0.301	0.350	0.468	0.501
1.3	0.3	0.169	0.185	0.254	0.306	0.378
	0.5	0.315	0.364	0.463	0.522	0.565
	0.7	0.612	0.743	0.828	0.886	0.944
1.5	0.3	0.316	0.417	0.488	0.542	0.608
	0.5	0.505	0.571	0.680	0.782	0.839
	0.7	0.823	0.909	0.950	0.966	0.995
1.7	0.3	0.448	0.561	0.654	0.734	0.797
	0.5	0.646	0.782	0.871	0.919	0.949
	0.7	0.901	0.952	0.988	0.998	0.998
2.0	0.3	0.658	0.745	0.829	0.896	0.928
	0.5	0.807	0.910	0.955	0.982	0.989
	0.7	0.957	0.985	0.999	0.997	0.999

^a Effect size is defined by the hazard ratio, with significance level 0.05.
^b Two years of accrual and two additional years of follow-up are assumed.
^c Correlation (ρ) is assumed to be 0.3, 0.5 or 0.7 between paired failure times.
Source: Mick R, Crowley JJ, and Carroll RJ, *Controlled Clin. Trials* 21:343–359, 2000, with permission from Elsevier Publishing.

8.4.2 MULTI-ARM DESIGNS

8.4.2.1 Randomized Selection Designs

In some cases, the aim of a phase II trial is not to identify whether an agent has biologic activity but instead to determine which agent of several should be chosen to be tested in the phase III setting. This is known as a selection or screening trial. The intent is not to definitively compare each of the regimens with one another but to pick the most promising agent to carry forward. The endpoint can be based on response rates, but survival or progression-free survival also is typically used as the criterion for picking the best arm. Selection designs are covered in detail in the chapter by P.Y. Liu et al. in this volume.

8.4.2.2 Comparison of Agent with Placebo

A randomized placebo controlled phase II design can be conducted successfully only if patients are willing to accept randomization to placebo. This could be in the setting where supportive care is the standard or in a group of patients who are stable after initial therapy or where there is no effective treatment. Korn et al.¹⁴ recommend a moderate one-sided alpha of 0.10 or 0.20 that reduces the required sample size. At the conclusion of the trial, a $p < 0.20$ would be considered evidence to carry the agent on to phase III testing. However this leads to a false positive conclusion one out of five times, and conducting a randomized design in the phase II setting with a placebo control arm establishes some legitimacy for comparison between the arms. Hence, there is a great temptation to interpret the results literally and not carry the agent forward to proper phase III testing.

Liu et al.¹⁸ calculated the probabilities of observing a hazard ratio greater than 1.3, 1.5, and 1.7 when the true hazard ratio is 1 between all treatments in a randomized phase II trial. In a two-arm trial with 40 patients per arm, the probabilities were 0.37, 0.17, and 0.07 for detecting the respective hazard ratios. Thus, the false-positive rates are very high if one treats randomized phase II trial results as conclusive.

8.4.2.3 Randomization Discontinuation Design

There can be substantial heterogeneity of tumor growth rates in patient populations. Some patients' tumors will grow slowly naturally. In order to distinguish antiproliferative activity of a novel agent from indolent disease, Rosner et al. proposed what they call a randomized discontinuation design (RDT).¹⁹ A generic study schema can be seen in [Figure 8.2](#). All patients are initially treated with the experimental agent (part I of trial), and these patients can be thought of as coming from a target population of patients with a given disease and stage. Patients without progression are randomized in a double-blind fashion to continuing therapy or placebo (part II). Patients who are noncompliant or experience adverse events are also typically not randomized. This allows the investigators to assess if apparent slow tumor growth is attributable to the drug or to the selection of patients with slow-growing tumors. By selecting a more homogeneous population, the randomized portion of the study may require fewer patients than would a study randomizing all patients.

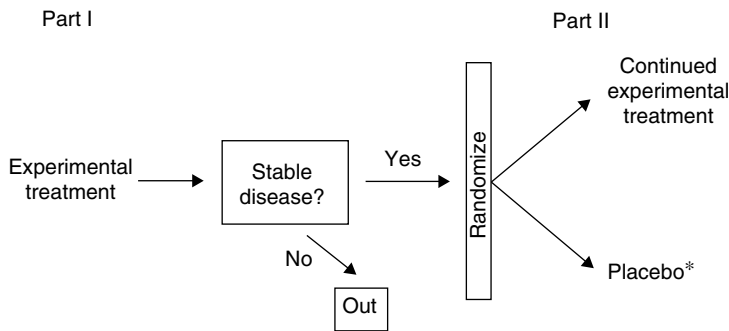


FIGURE 8.2 Schema for a randomised discontinuation design study.

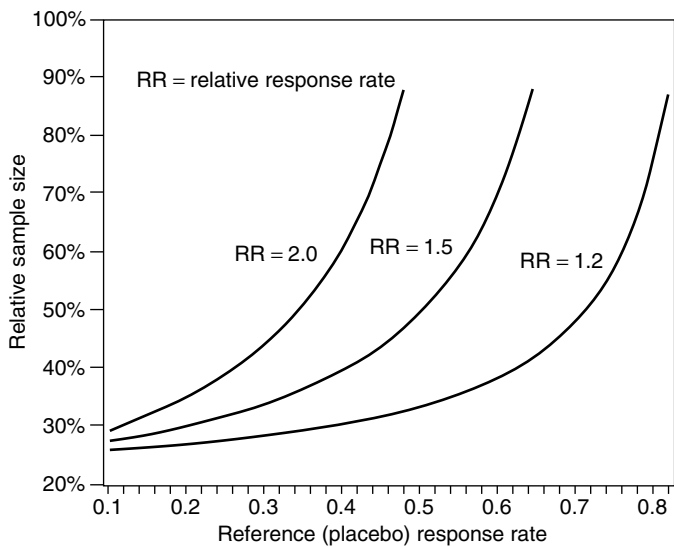


FIGURE 8.3 Sample size required in an RDT relative to that in a classic RCT (in %) for three different relative response rates in the source population: 1.2, 1.5, and 2.0. The frequencies of noncompliance and adverse reactions are assumed to be zero, and the sensitivity and specificity of identifying responders are assumed to be 0.8. (Reprinted from Kopec JA, Abrahamowicz M, and Esdaile JM, *J. Clin. Epidemiol.* 46(9):959–971, 1993, with permission from Elsevier publishing.)

Kopec et al.²⁰ have reviewed the advantages and limitations of discontinuation studies, and compared the RDT design to the classic randomized clinical trial (RCT) design in terms of clinical utility and efficiency (sample size). The response rate in the placebo group will be $r_0 = R_0/R$ where R_0 and R are the response rates on placebo and

active treatment in the source population. If there is perfect agreement in part I and part II of the trial, then the response rate r in the treatment group in part II would equal 1. However, that is not a reasonable expectation. The expected absolute effect of treatment in part II can be calculated as:

$$rd = r - r_0 = (R - R_0) \times \frac{SE_r}{(SE_r)(R) + (1 - SP_r)(1 - R)} \text{ where } SE_r \text{ and } SP_r \text{ are the}$$

sensitivity and specificity of identifying the responders in part I of the trial.

Figure 8.3 shows the relative sample size required for an RDT relative to a classic RCT when the placebo response rate and the relative response rate (R/R_0) vary. In this figure, noncompliance and adverse events are assumed to be zero. One sees the greatest gain in efficiency when the placebo response rate is low and the relative response rate is modest.

Figure 8.4 provides the relative sample size required for a RDT relative to a classic RCT when the sensitivity and specificity of identifying responders is varied. The relative response rate is fixed at 1.5. Improving the specificity of criteria for identifying responders may significantly reduce the relative sample size. Sensitivity has much less effect.

Kopec et al.²⁰ concluded that the RDT design is quite useful for studying the effect of long-term, noncurative therapies when the definition of clinically important effect is relatively small and that the use of a placebo should be minimized for

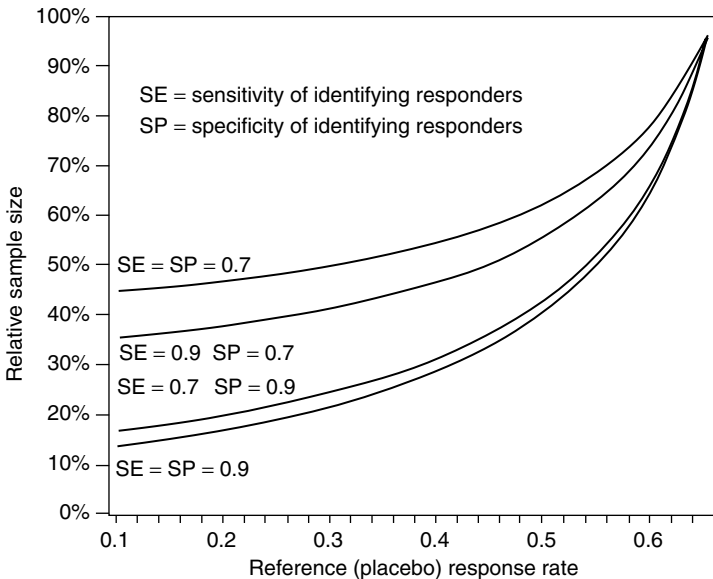


FIGURE 8.4 Sample size required in an RDT relative to that in a classic RCT (in %) for different sensitivities and specificities of identifying responders. The relative response rate is assumed to be 1.5, and the frequencies of noncompliance and adverse reactions are assumed to be zero. (Reprinted from Kopec JA, Abrahamowicz M, and Esdaile JM, *J. Clin. Epidemiol.* 46(9):959–971, 1993, with permission from Elsevier publishing.)

ethical or feasibility reasons. Conversely, the RDT design is limited if the objective of the study is to estimate the treatment effect and toxicity within the target population of patients with the disease of interest or if the treatment is potentially curative. The relative efficiency of the RDT design depends on the accuracy of the selection criteria with respect to identifying true treatment responders and to some degree those with good compliance and lack of limiting toxicities. As pointed out by Friedman et al.,²¹ because the RDT evaluates a highly selected sample, this design can overestimate benefit and underestimate toxicity. The RDT design, which can end up requiring a fairly large sample size, may be answering an irrelevant hypothesis; namely, given that one initially responds to a new agent, is more of the drug better than stopping at the time of response? It is easy to see how an agent initially could induce a high level of response but not look better than placebo with the administration of subsequent treatment. The converse may also be true. Because one concludes that a new agent is better than placebo in an RDT, it still might not have a level of activity that is desired for the entire target population.

8.5 CONCLUDING REMARKS

With the development of cytostatic agents, attention has been given to developing new phase II trial designs to address the expected lack of cytoreductive activity with these regimens. Endpoints that incorporate information about progression or survival are reasonable choices that can be straightforward to conduct. Interestingly, these same designs should be considered when planning a study with a cytoreductive agent. In many disease settings, response rates have not been found to be correlated with survival, so incorporating time-to-event endpoints as the primary objective may more accurately reflect an agent's true level of biologic activity in a particular disease setting.

REFERENCES

1. Forastiere AA, Metch B, Schuller DE, Ensley JF, Hutchins LF, Triozzi P, Kish JA, McClure S, VonFeldt E, and Williamson SK. Randomized comparison of cisplatin plus fluorouracil and carboplatin plus fluorouracil versus methotrexate in advanced squamous-cell carcinoma of the head and neck: a Southwest Oncology Group study. *J. Clin. Onc.* 10:1245–1251, 1992.
2. Mantel N and Byar DP. Evaluation of response-time data involving transient states: an illustration using heart transplant data. *JASA* 69:81–86, 1974.
3. Anderson JR, Cain KC, and Gelber RD. Analysis of survival by tumor response. *J. Clin. Onc.* 1(11):710–719, 1983.
4. Durie BGM, Jacobson J, Barlogie B, and Crowley J. Magnitude of response with myeloma frontline therapy does not predict outcome: importance of time to progression in Southwest Oncology Group chemotherapy trials. *J. Clin. Onc.* 22:1857–1863, 2004.
5. McHugh K and Kao S. Response evaluation criteria in solid tumors (RECIST): problems and need for modifications in paediatric oncology? *Brit. J. Radiol.* 76(907):433–436, 2003.

6. Erasmus JJ, Gladish GW, Broemeling L, Sabloff BS, Truong MT, Herbst RS, and Munden RF. Interobserver and intraobserver variability in measurement of non-small-cell carcinoma of the lung lesions: implications for assessment of tumor response. *J. Clin. Onc.* 21(13):2574–2582, 2003.
7. Kimura M and Tominaga T. Outstanding problems with response evaluation criteria in solid tumors (RECIST) in breast cancer. *Breast Cancer* 9(2):153–159, 2002.
8. Monetti F, Casanova S, Grasso A, Cafferata MA, Ardizzoni A, and Neumaier CE. Inadequacy of the new Response Evaluation Criteria in Solid Tumors (RECIST) in patients with malignant pleural mesothelioma: report of four cases. *Lung Cancer* 43(1):71–74, 2004.
9. Folkman J. Fighting cancer by attacking its blood supply. *Sci. Am.* 1996;275:150–154.
10. Markham M. Challenges associated with evaluating the clinical utility of noncytotoxic pharmaceutical agents in oncology. *J. Cancer Res. Clin. Oncol.* 1997;123:581–582.
11. Fleming, TR. One sample multiple testing procedures for phase II clinical trials. *Biometrics* 38:143–151, 1982.
12. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clin. Trials* 10:1–10, 1989.
13. Herndon, JE. A design alternative for two-stage phase II, multicenter cancer clinical trials. *Controlled Clin. Trials* 19:440–450, 1998.
14. Korn EL, Arbuck SG, Pluda JM, Simon R, Kaplan RS, and Christian MC. Clinical trial designs for cytostatic agents: are new approaches needed? *J. Clin. Oncol.* 19(1):265–272, 2001.
15. Mick R, Crowley JJ, and Carroll RJ. Phase II clinical trial design for noncytotoxic anticancer agents for which time to disease progression is the primary endpoint. *Controlled Clin. Trials* 21:343–359, 2000.
16. Von Hoff DD. There are no bad anticancer agents, only bad clinical trial designs. *Clin. Cancer Res.* 4:1079–1086, 1998.
17. Kalbfleisch JD and Prentice RL. *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons, Inc., 1980.
18. Liu PY, LeBlanc M, and Desai M. False positive rates of randomized phase II designs. *Controlled Clin. Trials* 20:343–352, 1999.
19. Rosner GL, Stadler W, and Ratain MJ. Randomized discontinuation design: application to cytostatic antineoplastic agents. *JCO* 20(22):4478–4484, 2002.
20. Kopec JA, Abrahamowicz M, and Esdaile JM. Randomized discontinuation trials: utility and efficacy. *J. Clin. Epidemiol.* 46(9):959–971, 1993.
21. Friedman LM, Furberg, CD, and DeMets DL. *Fundamentals of Clinical Trials*. 2nd ed. Littleton, MA: PSG Publishing Company, 1985.

9 Phase II Selection Designs

*P.Y. Liu, James Moon, and
Michael LeBlanc*

CONTENTS

9.1	Basic Concept.....	155
9.2	Sample Size Requirements.....	156
9.2.1	Binary Outcomes.....	156
9.2.2	Survival Outcomes.....	157
9.3	Variations of the Design.....	157
9.3.1	Designs with Minimum Activity Requirements.....	157
9.3.1.1	Binary Outcomes.....	159
9.3.1.2	Survival Outcomes.....	161
9.3.2	Designs with Minimum Advantage Requirements.....	162
9.3.3	Designs for Ordered Treatments.....	162
9.4	Concluding Remarks.....	163
	References.....	163

9.1 BASIC CONCEPT

When there are multiple promising new therapies in a disease setting, it may not be feasible to test all of them against the standard treatment in a definitive phase III trial. The sample sizes required for a phase III study with more than three arms could be prohibitive.¹ In addition, the analysis can be highly complex and prone to errors due to the large number of possible comparisons in a multi-arm study. An alternative strategy is to screen the new therapies first in a phase II setting and choose one to test against a standard treatment in a simple two-arm phase III trial. Selection designs can be used in such circumstances.

Simon et al.² first introduced statistical methods for ranking and selection to the oncology literature. In a selection design, patients are randomized to treatments involving new combinations or schedules of known active agents or new agents for which activity against the disease in question has already been demonstrated in some setting. In other words, the regimens under testing have already shown promise. Now the aim is to narrow the choice for formal comparisons to the standard therapy. With this approach one always selects the observed best treatment for further study, however small the advantage over the others may appear to be. Hypothesis tests are not performed. Sample size requirements are established so that, should a superior treatment

exist, it will be selected with a high probability. The necessary sample sizes are usually similar to those associated with pilot phase II trials before phase III testing.

Before proceeding further, it is important to note that although the statistical principles for selection designs are simple, their proper application can be slippery. Falsely justified by the randomized treatment assignment, a major pitfall of the selection design is to treat the observed ranking as conclusive and forego the required phase III testing. This practice is especially dangerous when a control arm is included as the basis for selection or when all treatment arms are experimental but a standard treatment does not exist for the particular disease. A treatment of choice can be declared with false justifications in these situations. If such a conclusion is desired, phase III studies with appropriately planned type I and type II errors should be conducted. Regarding a two-arm selection design, Sargent and Goldberg³⁻⁵ state, “The goal of the randomized phase II trial is to ensure that if one treatment is clearly inferior to the other, there is a small probability that the inferior treatment will be carried forward to a phase III trial.” Because of the design’s moderate sample sizes and lack of type I error control for false positive findings, the results are error prone when treated as ends in themselves.⁶ Prior to embarking on a selection study, it is vital that the investigators understand the design’s limitations and the proper interpretation of upcoming results.

9.2 SAMPLE SIZE REQUIREMENTS

9.2.1 BINARY OUTCOMES

Table 9.1 is reproduced from Simon et al.² for binary outcomes with $K = 2, 3,$ and 4 groups. The sample sizes were presumably derived by normal approximations to binomial distributions. With the listed N per group and true response rates, the correct selection probability should be approximately 0.90. A check by exact probabilities indicates that the actual correct selection probability ranges from 0.88 in most cases down to 0.86 when N is small. Increasing the sample size per group by 6 raises

TABLE 9.1
Sample Size per Treatment for Binary Outcomes and
0.90 Correct Selection Probability

Response Rates		N per Group		
$P_1, \dots, P_{K-1}$	P_K	$K = 2$	$K = 3$	$K = 4$
10%	25%	21	31	37
20%	35%	29	44	52
30%	45%	35	52	62
40%	55%	37	55	67
50%	65%	36	54	65
60%	75%	32	49	59
70%	85%	26	39	47
80%	95%	16	24	29

Source: Simon R, Wittes et al., *Cancer Treatment Reports*, 1985;69: 1375–1381.

the correct selection probability to 0.90 in all cases and may be worth considering when N is less than 30.

Except in extreme cases, Table 9.1 indicates the sample size to be relatively insensitive to baseline response rates (i.e., response rates of groups 1 through $K-1$). Since precise knowledge of the baseline rates is often not available, a conservative approach is to use always the largest sample size for each K , i.e., 37, 55, and 67 patients per group for $K = 2, 3$, and 4, respectively. While a total N of 74 for two groups is in line with large phase II studies, the total number of patients required for four groups, i.e., close to 270, could render the design impractical for many applications. Obviously the sample size can be reduced for differences greater than the 15% used for Table 9.1. However, if tumor response rate is the outcome of interest, it is generally low (e.g., 10%–20%) for many types of cancer, and an absolute 15% increase would certainly indicate a superior response rate. Such a treatment should be not missed in the selection process by inadequate sample sizes. Similarly, a correct selection probability of 0.90 also should be treated as the standard because a lower probability would result in too many false negative trials.

9.2.2 SURVIVAL OUTCOMES

For censored survival data, Liu et al.⁷ suggested fitting the Cox proportional hazards model, $h(t, z) = h_0(t) \exp(\beta'z)$, to the data where z is the $(K-1)$ dimensional vector of treatment group indicators and $\beta = (\beta_1, \dots, \beta_{K-1})$ is the vector of log hazard ratios. We proposed selecting the treatment with the smallest $\hat{\beta}_i$ (where $\hat{\beta}_K \equiv 0$) for further testing. Sample sizes for 0.90 correct selection probability were calculated based on the asymptotic normality of the $\hat{\beta}$. The requirements for exponential survival and uniform censoring are reproduced in Table 9.2. Simulation studies of robustness of the proportional hazards assumption found the correct selection probabilities to be above 0.80 for moderate departures from the assumption.

As with binary outcomes, the sample sizes become less practical when there are more than three groups or the hazard ratio between the worst and the best groups is smaller than 1.5.

Table 9.2 covers scenarios where the patient enrollment period is similar to the median survival of the worst groups. It does not encompass situations where these two quantities are quite different. Because the effective sample size for exponential survival distributions is the number of uncensored observations, the actual numbers of expected events are the same for the different rows in Table 9.2. For a 0.90 correct selection probability, Table 9.3 gives the approximate number of events needed per group for the worst groups. With $\int IdF$ as the proportion of censored observations, where I and F are the respective cumulative distribution functions for censoring and survival times, readers may find the expected event count more flexible for planning purposes.

9.3 VARIATIONS OF THE DESIGN

9.3.1 DESIGNS WITH MINIMUM ACTIVITY REQUIREMENTS

Though the selection design is most appropriate when adequate therapeutic effect is no longer in question, the idea of selection is sometimes applied to randomized

TABLE 9.2
Sample Size per Treatment for Exponential Survival Outcomes with One Year
Accrual and 0.90 Correct Selection Probability. Median = Median Survival in
Years for Groups 1 through $K-1$; Follow = Additional Follow-up in Years after
Accrual Completion; HR = Hazard Ratio of Groups 1 through $K-1$ vs. Group K

Median	Follow	HR =	K = 2			K = 3			K = 4		
			1.3	1.4	1.5	1.3	1.4	1.5	1.3	1.4	1.5
0.5	0		115	72	51	171	107	76	206	128	91
	0.5		71	44	31	106	66	46	127	79	56
	1		59	36	26	88	54	38	106	65	46
0.75	0		153	96	69	229	143	102	275	172	122
	0.5		89	56	40	133	83	59	160	100	70
	1		70	44	31	104	65	46	125	78	55
1	0		192	121	87	287	180	128	345	216	153
	0.5		108	68	48	161	101	72	194	121	86
	1		82	51	36	122	76	54	147	92	65

Source: Liu PY, et al., *Biometrics*, 1993;49:391–398.

TABLE 9.3
Expected Event Count per Group for the Worst Groups for Exponential
Survival and 0.90 Correct Selection Probability. HR = Hazard Ratio of Groups
1 through $K-1$ vs. Group K

K	HR		
	1.3	1.4	1.5
2	54	34	24
3	80	50	36
4	96	60	43

phase II trials when anticancer activities have not been previously established for the treatments involved. Alternatively, the side effects of the treatments could be substantial so that a certain activity level must be met in order to justify the therapy. In such cases, each treatment arm is designed as a stand-alone phase II trial with the same acceptance criterion for all arms. When more than one treatment arms are accepted, the observed best arm is selected for further study. The design typically specifies a null activity level that does not justify the further pursuit of a treatment and an alternative activity level that would definitely render a treatment worthy of more investigation. The sample size and an end-of-study acceptance criterion signifying rejection of the null hypothesis are then specified, with the one-sided type I error rate and power often set at 0.05 and 0.90, respectively. When there are K arms with selection as the end goal, we recommend Bonferoni adjustment of the individual arm’s type I error rate in order to maintain the study-wide type I error at 0.05. In

the following sections we examine the correct selection probability from these designs.

9.3.1.1 Binary Outcomes

For binary outcomes, standard designs without a selection component have been well developed; see the overview of phase II designs by Green in this volume. Table 9.4 lists three such designs. Designs B1 and B2 represent typical situations in new agent testing for cancer treatment where moderate tumor response rates in the 20% to 30% range already warrant further investigations. Design B3 is more for theoretical interest when response rates are near 50%.

Assume the same true response rate configuration as in Table 9.1, i.e., $P_1 = \dots = P_{K-1} < P_K$. Table 9.5 indicates that when the true P_1 and P_K values are the same as the null and alternative design parameters, respectively, i.e., 5%/20% for Design B1, 10%/30% for Design B2, and 40%/60% for Design B3, the chance of a correct selection result is approximately the same as the design power, 0.90, in all three cases. In other words, the operating characteristics of the phase II design dominate in this case, and the chance of an inferior arm's passing the acceptance level and further surpassing the best arm is negligible.

When the true P_1 and P_K are far higher than the null and alternative levels of the phase II design, all arms will meet the acceptance requirement, and selection design properties take over. When $P_K - P_1 = 15\%$, Table 9.1 results can be used as a general guide for correct selection probability. For example, when $K = 2$, the per arm sample sizes of 45, 36, and 62 in Table 9.4 compare favorably with the highest N of 37 from Table 9.1; therefore, the correct selection probabilities are generally 0.88 or higher when the minimum acceptance level is easily met by the best arm. However,

TABLE 9.4
Three Phase II Designs for Binary Data

K	Nominal α per Arm	N per Arm	Acceptance Level*	Exact α	Exact Power
Design B1: null level = 5%, alternative level = 20%					
2	0.025	45	$\geq 6/45$	0.0239	0.91
3	0.0167	50	$\geq 7/50$	0.0118	0.90
4	0.0125	50	$\geq 7/50$	0.0118	0.90
Design B2: null level = 10%, alternative level = 30%					
2	0.025	36	$\geq 8/36$	0.0235	0.89
3	0.0167	40	$\geq 9/40$	0.0155	0.89
4	0.0125	45	$\geq 10/45$	0.0120	0.91
Design B3: null level = 40%, alternative level = 60%					
2	0.025	62	$\geq 33/62$	0.0239	0.89
3	0.0167	69	$\geq 37/69$	0.0151	0.89
4	0.0125	75	$\geq 40/75$	0.0133	0.90

*Response rates signifying treatment worthy of further investigation.

TABLE 9.5
Correct Selection Probabilities for Binary Data Designs with Minimum Acceptance Level^a (3000 Simulations)

Design B1, testing 5% vs. 20%, true $P_K - P_1 = 15\%$					
True P_1/P_K^b	5%/20%	10%/25%	15%/30%	20%/35%	30%/45%
$K = 2$	0.91	0.96	0.95	0.94	0.91
$K = 3$	0.89	0.95	0.92	0.90	0.87
$K = 4$	0.90	0.92	0.90	0.88	0.84
Design B2, testing 10% vs. 30%, true $P_K - P_1 = 15\%$					
True P_1/P_K	10%/25%	15%/30%	20%/35%	30%/45%	40%/55%
$K = 2$	0.71	0.86	0.90	0.89	0.88
$K = 3$	0.69	0.85	0.87	0.83	0.82
$K = 4$	0.71	0.85	0.85	0.80	0.80
Design B2, testing 10% vs. 30%, true $P_K - P_1 = 20\%$					
True P_1/P_K	10%/30%	15%/35%	20%/40%	30%/50%	40%/60%
$K = 2$	0.89	0.95	0.96	0.95	0.94
$K = 3$	0.88	0.94	0.94	0.91	0.93
$K = 4$	0.90	0.95	0.93	0.93	0.92
Design B3, testing 40% vs. 60%, true $P_K - P_1 = 15\%$					
True P_1/P_K	40%/55%	45%/60%	50%/65%	55%/70%	60%/75%
$K = 2$	0.66	0.88	0.95	0.95	0.96
$K = 3$	0.62	0.86	0.92	0.92	0.94
$K = 4$	0.65	0.88	0.90	0.93	0.93
Design B3, testing 40% vs. 60%, true $P_K - P_1 = 20\%$					
True P_1/P_K	40%/60%	45%/65%	50%/70%	55%/75%	60%/80%
$K = 2$	0.88	0.97	0.98	0.99	0.99
$K = 3$	0.88	0.96	0.98	0.98	0.99
$K = 4$	0.90	0.97	0.98	0.99	0.99

^a Probability of the true best arm passing the acceptance level and being the observed best arm.

^b $P_1 = \dots = P_{K-1}$.

when $P_1 = 30\%$, $P_K = 45\%$, and $K = 4$, per Table 9.1, the N required per arm is 62 for an approximate 0.90 correct selection probability. The corresponding N for $K = 4$ in Designs B1 and B2 are 50 and 45, respectively; therefore, the correct selection probabilities in Table 9.5 are less than 0.90. When $P_K - P_1 = 20\%$, as is the distance between the alternative and null values for Designs B2 and B3, correct selection probabilities are approximately 0.90 or higher in all cases examined.

In general, applying Bonferoni adjustment to per-arm type I errors performs adequately with respect to correct selection probabilities in the situations examined, of which the parameter ranges cover most cancer trial applications. The approach is appropriate when the emphasis is on the initial step of screening the treatments for minimum acceptable anti-tumor activities. If meeting the minimum activity level is relatively assured and the emphasis is on selection, it would be more appropriate to design the trial with the larger sample size between what is required for the phase II and the selection portions.

9.3.1.2 Survival Outcomes

The approach for binary outcomes can be applied to survival outcomes as well. With Bonferoni adjustment for type I errors, phase II designs with null and alternative hypotheses are used first to test the minimum acceptance level. Selection ensues when two or more arms are accepted. Assuming exponential survival and uniform censoring, Table 9.6 lists correct selection probabilities for one scenario from Table 9.2. The results are generalizable to other accrual and follow-up length combinations because the event count, hazard ratio, and the number of arms, which are represented here, are the determinants for power or selection probability under the exponential assumption.

Similar to binary data results, when $m_1 = 0.75$ and $m_K/m_1 = 1.5$ as designed, the phase II operating characteristics dominate, and the correct selection probability is approximately 0.90 — the same as the individual arm's planned power. The correct selection probability is poor when $m_1 = 0.75$ and $m_K/m_1 = 1.3$. When the worst median is higher than the null level of 0.75, selection properties begin to apply. For example, for Design S3, $K = 4$, $N = 147$ per arm, and the worst median is 0.95, the expected exponential event count is 75 with 1 year accrual and an additional 0.5 year follow-up. Compared to Table 9.3, the event count required for 0.90 correct selection probability is 96, 60, and 43 for $m_K/m_1 = 1.3$, 1.4, and 1.5, respectively, thus explaining the corresponding correct selection probabilities of 0.86, 0.93, and 0.97 in Table 9.6.

TABLE 9.6
Correct Selection Probabilities for Exponential Survival Design^a
with 1 Year Accrual, 0.5 Year Follow-Up, and Minimum
Acceptance Level (3000 Simulations)

True m_1	Correct Selection Probability for $m_K/m_1^b =$		
	1.3	1.4	1.5
Design S1: $K = 2$, $N = 124$ per arm, observed median acceptable if >0.95			
0.75	0.57	0.78	0.91
0.85	0.84	0.94	0.97
0.95	0.90	0.96	0.98
Design S2: $K = 3$, $N = 137$ per arm, observed median acceptable if >0.96			
0.75	0.53	0.77	0.89
0.85	0.84	0.93	0.97
0.95	0.88	0.95	0.98
Design S3: $K = 4$, $N = 147$ per arm, observed median acceptable if >0.96			
0.75	0.54	0.77	0.90
0.85	0.83	0.93	0.97
0.95	0.86	0.93	0.97

^a All designs are for detecting a hazard ratio of 1.5 over the null median of 0.75 with one-sided $0.05/K$ type I error and 0.90 power for each arm.

^b $m_1 = \dots = m_{K-1} < m_K$.

9.3.2 DESIGNS WITH MINIMUM ADVANTAGE REQUIREMENTS

Some authors propose changing the selection criterion so that the observed best treatment will be further studied only when its minimum advantage over all other treatments is greater than some positive Δ ; otherwise the selection will be based on other factors. Table 9.7 gives some binary data sample size requirements³⁻⁵ for $\Delta = 0.05$.

While this approach is appealing because the decision rule is easier to carry out in practice, the sample sizes required generally more than double those in Table 9.1. For example, for $K = 2$, $P_1 = 35\%$, $P_2 = 50\%$, and 0.90 correct selection probability, it can be interpolated from Table 9.1 that 36 patients per treatment are required for $\Delta = 0$. With the same configuration, the required number of patients is 75 per group for $\Delta = 5\%$. Clearly when $\Delta > 5\%$ the sample size requirement would be impractical. Even with $\Delta = 5\%$ and 75 patients per group, the results are by no means definitive when a greater than 5% difference is seen. When $P_1 = P_2 = 35\%$ and $N = 75$, the chance of observing $|p_1 - p_2| > 5\%$ is approximately 0.52. On the other hand, with $\Delta = 0$ and $N = 36$ per Table 9.1, the chance of observing $p_2 - p_1 > 5\%$ is approximately 0.81 when the true $P_1 = 35\%$ and $P_2 = 50\%$. Therefore, with an incremental gain in the probability of correctly observing a $\Delta > 5\%$ but double the sample size, this approach may only be practical when patient resources are plentiful. While precision is improved with the larger sample sizes, the results are nevertheless nondefinitive.

9.3.3 DESIGNS FOR ORDERED TREATMENTS

When the $K (\geq 3)$ treatments under consideration consist of increasing dose schedules of the same agents, the design can take advantage of this inherent order. A simple method is to fit regression models to the outcomes with treatment groups coded in an ordered manner. Logistic regression for binary outcomes and the Cox model for survival are obvious choices. A single independent variable with equally spaced scores for the treatments could be included in the regression. If the sign of the observed slope is in the expected direction, the highest dose with acceptable toxicity is selected for further study. Otherwise the lowest dose schedule would be selected.

TABLE 9.7
Sample Size per Treatment for Binary Outcomes and
0.90 Correct Selection Probability when Requiring
an Absolute 5% Minimum Advantage for Selection

Response Rates		N per Group		
$P_1, \dots, P_{K-1}$	P_K	K=2	K=3	K=4
5%	20%	32	39	53
15%	30%	53	77	95
25%	40%	71	98	119
35%	50%	75	115	147

Source: Sargent DJ, Goldberg RM, *Statistics in Medicine*, 2001; 20:1051–1060. Sargent DJ, *Statistics in Medicine*, 2002; 21:628.

Compared to the nonordered design, this approach should require smaller sample sizes for the same correct selection probability. Limited simulations were conducted with the following results. For binary data with $K = 3$, $P_1 = 40\%$, $P_3 = 55\%$, $P_1 \leq P_2 \leq P_3$, approximately $N = 35$ per arm is needed for a 0.90 chance that the slope from the logistic regression is positive. Compared to $N = 55$ given in [Table 9.1](#), this is a substantial reduction in sample size. Similarly, for $K = 4$, $P_1 = 40\%$, $P_4 = 55\%$, $P_1 \leq P_2 \leq P_3 \leq P_4$, 40 patients per arm are needed instead of 67. For exponential survival data with a 1.5 hazard ratio between the worst groups and the best group, approximately 28 and 32 events per group are needed for the worst groups for $K = 3$ and 4, respectively, as compared to 36 and 43 given in [Table 9.3](#).

9.4 CONCLUDING REMARKS

The statistical principles of selection design are simple and adaptable to various situations in cancer clinical research. Applied correctly, the design can serve a useful function in the long and arduous process of new treatment discovery. However, as mentioned in the beginning, the principal misuse of the design is to treat the results as ends in themselves without the required phase III investigations. We previously published the false positive rates of this misapplication.⁶ It was shown that impressive looking differences arise with high frequencies purely by chance with selection design sample sizes. We also pointed out that performing hypothesis tests post hoc changes the purpose of the design. If the goal is to reach definitive answers, then a phase III comparison should be designed with appropriate analyses and error rates. Testing hypotheses with selection sample sizes can be likened to conducting the initial interim analysis for phase III trials. It is well known that small sample type I error assessments are unstable and extremely stringent p values are required to stop the trial at this early stage.

Lastly, when historical benchmarks are not available for setting a minimum acceptance activity level, the inclusion of a standard or control treatment in a selection design should be utilized with caution. Without a control arm, any comparison between the current standard and the observed best treatment from a selection trial is recognized as informal because the limitations of historical comparisons are widely accepted. When a control arm is included for randomization, the legitimacy for comparison is established and there can be great temptation to interpret the results literally and move on. If there are no efficacy differences between treatments, the chance of observing an experimental treatment better than the control is $(K - 1)/K$, i.e., $1/2$ for $K = 2$, $2/3$ for $K = 3$, etc. Again, an observed advantage for an experimental treatment simply means its substantial inferiority is unlikely so that further testing may be warranted and must be conducted for definitive comparisons.

REFERENCES

1. Liu PY and Dahlberg S. Design and analysis of multiarm clinical trials with survival endpoints. *Controlled Clinical Trials* 1995;16:119–130.
2. Simon R, Wittes RE, and Ellenberg SS. Randomized phase II clinical trials. *Cancer Treatment Reports* 1985;69:1375–1381.
3. Sargent DJ and Goldberg RM. A flexible design for multiple armed screening trials. *Statistics in Medicine* 2001;20:1051–1060.

4. Lui KJ. Letter to the editor: A flexible design for multiple armed screening trials. *Statistics in Medicine* 2002;21:625–627.
5. Sargent, DJ. Author's reply to letter to the editor. *Statistics in Medicine* 2002;21:628.
6. Liu PY, LeBlanc M, and Desai M. False positive rates of randomized phase II designs. *Controlled Clinical Trials* 1999;20:343–352.
7. Liu PY, Dahlberg S, and Crowley J. Selection designs for pilot studies based on survival. *Biometrics* 1993;49:391–398.
8. Green S and Dahlberg S. Planned versus attained design in phase II clinical trials. *Statistics in Medicine* 1992;11:853–862.
9. Gibbons JD, Olkin I, and Sobel M. *Selecting and Ordering Populations: A New Statistical Methodology*. New York: Wiley, 1977.

Part III

Phase III Trials

10 On Use of Covariates in Randomization and Analysis of Clinical Trials

Garnet L. Anderson, Ph.D., Michael LeBlanc, Ph.D., P.Y. Liu, Ph.D., and John Crowley, Ph.D.

CONTENTS

10.1	Introduction	167
10.2	Randomization Schemes	168
10.2.1	Permuted Blocks Randomization.....	168
10.2.2	Pocock–Simon, or Minimization, Approach	169
10.2.3	Stratified Randomization	170
10.3	Basis for Inference	170
10.4	Analytic Approaches	171
10.5	Design of Simulation Studies	172
10.6	Results	173
10.7	Discussion	177
	Acknowledgment.....	177
	References	178

10.1 INTRODUCTION

Multiple strategies exist for using covariate information in randomized treatment allocation schemes. Since most clinical trials cannot be repeated, reliance on balance in expectation may be considered insufficient. Hence, “Block what you can and randomize what you cannot”¹ is the wisdom often invoked for clinical trial design. Kalish and Begg² review a comprehensive list of treatment allocation strategies. The most common and straightforward method for using covariates is stratified randomization, an algorithm that applies permuted blocks to individual strata.³ When the ratio of sample size to strata is small, however, stratified randomization may actually increase the imbalance for given covariates. For these settings, minimization methods that dynamically balance the treatment assignment marginally for each factor have been proposed.^{4–6}

Use of covariates in the analysis of randomized experiments often is guided by our understanding of linear models. In this setting, covariates are not needed to produce an unbiased estimate of effect or a test of appropriate size but may be used to improve power by reducing the residual variance. In nonlinear models, however, omitting covariates may reduce efficiency because the treatment effect estimators are biased toward zero.^{7–10}

Data analysis is made more complicated by randomization methods that use covariates.^{11–13} Green and Byar demonstrated that an unstratified analysis of binary data generated from a trial using important prognostic covariates in a stratified treatment allocation rule yields a conservative test with noticeably reduced power. A stratified analysis in this setting preserves the nominal size of a test.¹⁴ For more complex randomization strategies, such as the adaptive designs, however, there is no direct link between the covariate structure in the design and the test statistic.

The Southwest Oncology Group (SWOG), an NCI-funded cooperative group, conducts numerous phase III randomized trials of cancer therapeutics. These trials are designed to test the effects of specific treatment regimens on survival or relapse-free survival. Baseline information on disease status and patient characteristics is often used in randomization. SWOG has adopted a biased-coin Pocock–Simon randomization rule to assure balance on the margins of the covariates.⁴ In reporting results, the majority of these studies use stratified logrank statistics or Cox proportional hazards regression to incorporate covariates. It is of interest to know if there is a preferred analysis approach in these settings.

The purpose of this investigation is to characterize the performance of the most commonly used survival analysis-based tests when applied to trials employing randomization rules that differ in their use of covariates. Several factors come into play in these settings: the number of strata or covariates and their distribution in the sample, the magnitude of the covariate and treatment effects over time, and the degree of censoring. Though we cannot be comprehensive in examining these factors, we examined a variety of settings, some specifically chosen to violate model assumptions or create instability, with the hope that these would provide useful insight.

10.2 RANDOMIZATION SCHEMES

Allocation rules generally fall into three categories based on their use of covariates: (1) randomization rules that are independent of covariates; (2) rules that promote balance marginally for each covariate; and (3) rules that balance the treatment arms within each stratum. We selected one allocation rule from each category to compare.

10.2.1 PERMUTED BLOCKS RANDOMIZATION

The easiest and most popular randomization strategy that ignores covariates is permuted blocks (PB).^{15,16} In this approach, blocks of treatment assignments are generated with each treatment arm appearing in random order in the predetermined proportion, and treatments are then assigned sequentially to patients as they enter the trial. For the case of block size N , where N is the total sample size, a PB approach merely assures equal proportions in each arm. Typically smaller block sizes are used

1	A	A	B	B
2	A	B	A	B
3	A	B	B	A
4	B	A	B	A
5	B	A	A	B
6	B	B	A	A

FIGURE 10.1 Example of permuted blocks. Assume that 100 patients are to be randomized to two arms (A,B), using a block size of 4. There are 6 potential random blocks. One may randomly assign 100 patients to these treatment assignments by randomly selecting with replacement 25 blocks of assignments by rolling a six-sided die 25 times and then for each roll placing the corresponding blocks of assignments in the same sequence.

(e.g., four, six, and eight for two-arm trials) as illustrated in Figure 10.1. Since balance is attained at the completion of each block, these smaller block sizes provide the advantage of balancing on the implicit covariate of time of entry, an important feature if there is a significant possibility of drift in patient characteristics over time. In this article, however, we examine a PB design with block size N , a method that imposes the fewest constraints on the randomization.

10.2.2 POCOCK–SIMON, OR MINIMIZATION, APPROACH

The Pocock–Simon (PS),⁴ or minimization,⁶ approach is an adaptive allocation scheme that uses the covariate stream and a measure of covariate imbalance to determine treatment assignment. Imbalance may be defined as the sum of the treatment imbalance for each covariate, weighted by a measure of the perceived importance of the covariate if desired. If A_{ij} and B_{ij} represent the number of individuals currently assigned to arm A and Arm B, respectively, then a simple measure of imbalance would be $\Delta = \sum \text{abs}(A_{ij} - B_{ij})$ for covariate i value j . To emphasize balance for selected covariates, weights (w_{ij}) can be assigned to these covariates and incorporated into the $\Delta = \sum w_{ij} \text{abs}(A_{ij} - B_{ij})$. The measure of imbalance may also be defined for combinations of covariates to ensure balance within such strata, but this approach is most often employed to provide an equal distribution on each covariate univariately.

In adaptive randomization, treatment arms are assigned at random, usually with equal probability when there is perfect balance. When there is imbalance, the next assignment using the minimization approach is automatically determined to be the arm that produces the least imbalance. Under the biased coin generalization, the SWOG standard, the next assignment randomizes a patient to the arm that best reduces any imbalance with higher probability (e.g., ³/₄). Figure 10.2 illustrates how the next treatment assignment in a clinical trial would be determined when balance on two covariates, in this case age and sex, is sought. Hence the algorithm is designed to ensure approximately the same proportions of males and females on

	50–59		60–69		70–79		Total		
Sex	A	B	A	B	A	B	A	B	δ_{SEXj}
M	7	9	13	12	7	5	27	26	1
F	5	4	8	8	14	16	27	28	1
Total	12	13	21	20	21	21			
δ_{AGEj}	1		1		0				

FIGURE 10.2 Example of a simple adaptive randomization. Assume there are two treatment arms (A,B) and two stratification variables: Age (50–59, 60–69, 70–79) and Sex (M,F). For each possible value j of covariate i let $\delta_{ij} = \text{abs}(A_{ij} - B_{ij})$, the absolute value of the difference in the numbers currently assigned to A and B among patients with covariate i value j . Then define the measure of total imbalance to be $\Delta = \sum \delta_{ci}$. If the current allocation by age and sex is given as shown, then the present value of $\Delta = 1 + 1 + 1 + 1 + 0 = 4$. If the next individual is a male, age 60–69, the only δ values that would change are those for males and for 60–69 year olds. Assignment to A would result in $\Delta = 2 + 1 + 1 + 2 + 0 = 6$; assignment to B would result in $\Delta = 0 + 1 + 1 + 0 + 0 = 2$. Hence, B would be assigned with probability 3/4. If the next individual is a female, age 50–59, assignment to A would result in $\Delta = 1 + 0 + 0 + 1 + 0 = 2$, and assignment to B would give $\Delta = 1 + 2 + 2 + 1 + 0 = 6$. In this case A would be assigned with probability 3/4.

each arm and the same age distribution on each arm, but it may not necessarily give the same age distribution among males in both arms. We examined this biased coin procedure because it ensures reasonable covariate balance while avoiding the potential pitfalls of deterministic assignments.

10.2.3 STRATIFIED RANDOMIZATION

For a stratified randomization, we chose the stratified block (SB) algorithm that applies permuted blocks within each stratum. A stratified block design is easy to implement and common in multicenter studies, where blocking within centers is usually recommended when the sample size per center permits. Under well-controlled conditions, such as animal studies and agricultural studies, this approach best assures balance on the design covariates. In clinical trials where the sample size and overall distribution of covariates is unpredictable, balance may not be achieved because of incomplete blocks. A stratified permuted block design is illustrated in [Figure 10.3](#).

10.3 BASIS FOR INFERENCE

Randomized experiments afford two options for interpreting the results, randomization tests or population models. An advantage of randomization tests is that they are free from assumptions about the sampling frame. They lack some appeal in that, strictly speaking, the conclusions have a narrow applicability. If one were to use a randomization test as the basis for inference in a randomized trial, there is general

Age	Sex	Randomizations
50–59	M	A,B,B,A,A,B,A,B,B,A,A,B,A,B,A,A,B,A,A,B,B
50–59	F	B,A,A,B,A,B,A,B,B,A,A,B,A,B,A,A,B,B,A,B,A,B
60–69	M	A,B,A,B,B,B,A,A,A,B,A,B,A,B,B,A,B,A,A,B,B,A
60–69	F	A,A,B,B,A,B,A,B,A,A,B,B,B,A,A,B,A,B,A,B,B,A,A
70–79	M	B,B,A,A,A,A,B,B,A,B,B,A,B,A,B,B,A,A,A,A,B,B
70–79	F	A,B,B,A,A,B,B,A,B,A,B,A,B,A,A,A,A,B,B,A,B,B,A

FIGURE 10.3 An example of stratified permuted block randomization. Assume that 100 patients are to be randomized to two arms (A, B) with stratification by age (50–59, 60–69, 70–79) and sex (M, F). Using the permuted blocks identified in [Figure 10.1](#), one would generate six separate randomization lists, such as those shown. Unless one restricted the accrual to each of these strata to assure that recruitment ended at the completion of a block, in this case a multiple of four, perfect balance may not be achieved.

agreement that the structure of the allocation scheme must be incorporated in the analysis to preserve power. The implementation of a randomization test is straightforward for PB and SB randomizations. For nondeterministic adaptive designs, randomization tests can be obtained by comparing the observed test statistic to the distribution obtained from simulations that condition on the observed covariates, order of enrollment, and outcomes. Such an approach, though feasible, is cumbersome. We are not aware of anyone who has used this in published applications. For minimization designs, where a large fraction of the randomization assignments are strongly influenced or even dictated by the preceding assignments and the covariates of the current and previous subjects, the proper specification of a randomization test is not clear.

Inference based on population models assumes that the sample under test is representative of a larger population to which the results apply. This assumption does not strictly hold for most clinical trials in the sense that the study population usually represents volunteers rather than a probability sample of eligible patients. This approach is simpler to implement, however, in that it allows us to rely on the usual distribution theory for test statistics. Issues of generalizability are typically dealt with in the interpretation of results by considering eligibility and actual study subject characteristics. Because of its popularity, we evaluated this approach.

10.4 ANALYTIC APPROACHES

Analyses of survival data from clinical trials are typically based on either unstratified or stratified logrank tests or on a test for the treatment assignment variable from a proportional hazards model.¹⁷ Each of these can be developed in the context of a Cox proportional hazards (PH) model given by $h(t; x, z) = h_0(t; z_2) \exp(\beta x + \alpha z_1)$, where t is time from randomization, x represents a binary treatment assignment

variable, $z = (z_1, z_2)$ is a vector of covariates where z_1 corresponds to those covariates used in the regression function, z_2 represents the covariates used in stratification, and $h_0(t; z_2)$ is the baseline hazard function defined for each level of z_2 .

The logrank test, a commonly employed test statistic for comparing survival curves, can be obtained from a PH model as a maximum partial likelihood score test for $\beta = 0$ in the case where no other covariates are included.¹⁸ The natural generalization to a PH regression based test is the analogous score statistic, where covariates are used as regressors in the model with a single baseline hazard function h_0 . A fully stratified logrank test or generalized score statistic can be computed in a similar fashion, where a separate baseline hazard function for each stratum is allowed.

10.5 DESIGN OF SIMULATION STUDIES

We evaluated the impact of randomization schemes on the size and power of these analytic strategies in simulated clinical trials using both purely hypothetical trial scenarios as well as several derived from completed SWOG trials.

For each hypothetical trial, the basic scenario was a two-arm trial with 400 patients (200 per arm). The underlying survival models were derived from PH models assuming the existence of up to three binary covariates using the hazard models

$$h(t; x, z) = h_0(t) \exp(\beta x) \quad (\text{A})$$

$$h(t; x, z) = h_0(t) \exp\{\beta x + \ln(0.33)z_1 + \ln(1.5)z_2 + \ln(2)z_3\} \quad (\text{B})$$

$$h(t; x, z) = h_0(t; z_1, z_2, z_3) \exp(\beta x), \quad (\text{C})$$

where $h_0(t)$ represents the hazard function from the exponential distribution for models A and B. For all models, $x = 0, 1$ represents the randomization assignment, and z is the vector of binary covariates that jointly define membership into eight strata. For model C with nonproportional covariate effects, the eight baseline hazard functions $h_0(t; z_1, z_2, z_3)$ were defined by Weibull (λ_j, κ_j) functions $h(t; x) = h_0(t, \lambda_j, \kappa_j) \exp(\beta x)$, where $h_0(t, \lambda_j, \kappa_j) = \lambda_j \kappa_j (\lambda_j t)^{\kappa_j - 1}$, where λ_j and κ_j are the scale and shape parameters specified for each unique covariate combination j . The Weibull family of distributions was chosen because of the degree of flexibility it allows in describing nonproportional hazard functions. Values of (λ_j, κ_j) used were (0.2, 0.7), (0.1, 0.8), (0.6, 1), (0.1, 1.2), (0.2, 1.5), (0.5, 2), (0.2, 3). Note for $\kappa_j = 1$, the Weibull model reduces to the exponential (λ) distribution (constant hazard) and all covariate effects are proportional. When $\kappa > 1$ ($\kappa < 1$), the baseline hazard functions are decreasing (increasing) with time. Hazard functions associated with a covariate will be nonproportional when values of κ_j differ across levels of the covariate. To examine the setting of highly stratified allocation, model B was expanded to include five independent binary covariates (32 strata). Coefficients for these covariates in the PH model generating the survival endpoint were: $\ln(0.33)$, $\ln(1.5)$, $\ln(2)$, $\ln(0.67)$, $\ln(1.5)$.

For each simulated trial, patient level covariates were generated with all combinations having equal probability. Using these covariates and their associated effects, two potential survival times were generated, one for each potential treatment

assignment. Censoring times and the corresponding order of patient presentation were randomly generated from a uniform distribution to represent constant accrual rates over a designated interval. The two corresponding censored survival times and indicators were determined by comparing the survival times associated with each potential treatment assignment and with the censoring time for each patient.

Three treatment allocation rules (PB, PS, and SB) were applied to this same source data to simulate three trials in the same set of patients. As each patient in a trial was randomized according to the current rule, the appropriate censored survival times and indicators were selected from the source data to form the simulated trial data. All of the test statistics (Logrank, PH, Stratified Logrank) were then calculated for that trial. For each configuration, 5000 trials were simulated. Performance of each test statistic was assessed by calculating the proportion of test statistics exceeding 1.96 (corresponding to a 0.025-level, one-sided test) to estimate size (under $\beta = 0$) and power (under $\beta = \beta_A$).

To examine realistic settings with respect to covariate distribution and effects, we conducted a similar set of simulations motivated by five completed SWOG trials known by protocol numbers *SWOG-8516*, *SWOG-8501*, *SWOG-9210*, *SWOG-9308*, and *SWOG-8738*.^{19–23} Key features of these trials extracted for the simulations are shown in [Table 10.1](#). The number of design strata in these studies ranged from 4 to 38. For each stratum j where sample size permitted, we estimated the scale and shape parameters of a censored Weibull (λ_j , κ_j) distribution. These Weibull hazard functions then formed the basis of the survival times for patients generated under model C above.

The SWOG trial based simulations were conducted using the same general plan as described for the purely hypothetical trials while preserving as much of the underlying data structures of the original SWOG trials as possible. In particular, for each simulated SWOG trial, the sample size N was that of the original trial, and the covariates for the N patients were obtained by sampling with replacement from the observed covariate vectors. β_A was calculated separately for each trial to provide approximately 80% power for a one-sided 0.025-level test ($\ln(1.5)$ for protocols *SWOG-8516* and *SWOG-8501*, $\ln(1.65)$ for *SWOG-9210*, $\ln(1.4)$ for *SWOG-9308*, and $\ln(1.6)$ for *SWOG-8738*). Censoring was chosen to be uniform ($0, u_k$) and varied between studies, yielding 15% to 60% censoring depending on the survival distributions. The results of these simulations are summarized with estimated size and power as described above.

10.6 RESULTS

[Table 10.2](#) presents the estimated size and power for each combination of randomization schemes and test statistics in the purely hypothetical trial scenarios. Each combination provided a statistically valid test, as the estimated size was never significantly above the nominal 0.025 level even in the settings where some of the underlying assumptions are clearly violated (nonproportional covariate effects).

There is evidence of conservatism in the logrank statistic under randomization schemes that use covariate information. In each scenario where covariates were predictive of survival, the trials that used either the PS or SB design show estimated type

TABLE 10.1
Selected Parameters from five SWOG Studies Used as the Bases for Simulation Studies

Protocol #	Cancer Type/Site	N	No. of Covariates ^a	Strata	Weibull Parameters Pairs (λ , κ)
8516	Non-Hodgkins lymphoma	435	5/3	38	27 cells with (0.185, 0.845), 11 remaining cells having λ = (0.205, 0.125, 0.097, 0.118, 0.070, 0.207, 0.570, 0.155, 0.096, 0.222, 0.171) κ = (1.46, 0.915, 0.865, 1.38, 0.796, 0.615, 0.840, 0.573, 0.566, 0.652, 0.637)
8501	Ovary	546	4/2	29	20 cells with (0.192, 1.231), 9 remaining cells having λ = (0.110, 0.182, 0.267, 0.191, 0.284, 0.168, 0.070, 0.208, 0.255) κ = (0.943, 1.239, 1.366, 1.218, 1.38, 1.256, 0.89, 1.506, 1.283)
9210	Multiple myeloma	247	2/0	6	λ = (0.309, 0.309, 0.286, 0.262, 0.343, 0.344) κ = (1.30, 1.30, 1.33, 1.148, 1.454, 1.268)
9308	Non-small cell lung	414	2/2	4	λ = (1.029, .605, 1.643, 1.092) κ = (1.213, 1.209, 1.031, 1.151)
8738	Non-small cell lung	356	2/2	6	λ = (1.079, 2.248, 0.914, 1.379, 1.737, 1.471) κ = (1.086, 1.55, 0.903, 0.922, 1.483, 1.190)

^aNumber of covariates used in the randomization scheme/number found to be significant (p<0.05) in a Cox proportional hazards model analysis.

I errors rates less than 0.025 and in some cases as low as 0.005. Under the simple PB scheme, however, the size of the logrank statistic was unaffected by the presence of predictive covariates. The estimated type I error rates for the stratified logrank or proportional hazards based test appear unaffected by differences in randomization strategies.

Using covariates in the randomization appears to have a very modest effect on power throughout all of these configurations. Within any trial scenario and analytic approach, the differences in the estimated power between randomization strategies are less than 5%.

The effects of the analytic approach on power are more noticeable. The estimated power for the PH model and stratified logrank approaches was higher than for the logrank test in every case where there were important covariates, regardless of whether these covariates were used in treatment allocation. There was also a suggestion that the PH model yields increased power in the setting of many covariates with PH effects; in all other cases examined, the PH and stratified logrank approaches performed comparably, or the stratified logrank test yielded slightly better power.

TABLE 10.2

Estimated Size and Power^a for Each Combination of Randomization Rule and Analytic Approach Based on 5000 Simulated Trials Derived from Models (A)–(C)

Model & Allocation	Analysis Approach					
	Logrank		PH Model ^c		Stratified Logrank	
Rule ^b	Size	Power	Size	Power	Size	Power
<i>8 strata, no covariate effects</i>						
PB	0.023	0.923	0.026	0.923	0.026	0.902
PS	0.025	0.914	0.027	0.916	0.025	0.896
SB	0.025	0.920	0.025	0.918	0.025	0.912
<i>8 strata, moderate PH covariate effects</i>						
PB	0.025	0.755	0.027	0.904	0.022	0.889
PS	0.012	0.790	0.023	0.908	0.024	0.890
SB	0.009	0.790	0.023	0.905	0.023	0.889
<i>8 strata, large PH covariate effects</i>						
PB	0.025	0.579	0.025	0.930	0.026	0.915
PS	0.004	0.610	0.026	0.929	0.028	0.913
SB	0.005	0.584	0.029	0.926	0.026	0.926
<i>8 strata, non-PH covariate effects</i>						
PB	0.027	0.654	0.026	0.820	0.024	0.857
PS	0.013	0.676	0.025	0.824	0.027	0.855
SB	0.013	0.697	0.027	0.841	0.027	0.874
<i>32 strata, PH covariate effects</i>						
PB	0.027	0.764	0.027	0.905	0.026	0.825
PS	0.014	0.799	0.024	0.914	0.027	0.843
SB	0.015	0.792	0.027	0.908	0.026	0.853

Estimated type I error significantly less than the nominal 0.025

Estimated power significantly lower than the best performing approach

^a $2 \times \text{SE}(\text{estimated size}) = 0.004$, $2 \times \text{SE}(\text{estimated power}) = 0.011$.

^b PB: permuted block; PS: Pocock–Simon adaptive allocation; SB: stratified permuted block.

^c PH: proportional hazards.

Similar patterns are evident in the corresponding results for the five SWOG trials (Table 10.3). The estimated size did not significantly exceed the nominal level (0.025) in any of these simulated studies. In fact, the estimated type I error derived from actual trial covariate distributions, and their estimated effects on survival were more tightly clustered around 0.025 than was observed in the purely hypothetical studies. In the *SWOG-9308* trial, which exhibited the strongest covariate effects, the maximum range in type I errors was 0.019 to 0.028, still smaller than most of the purely hypothetical settings.

The estimated power for each combination of randomization scheme and analysis approach in the simulated SWOG trials also remained in a narrow window. The observed optimal configuration of randomization scheme and test statistic yielded at most a 5% increment in power over the least favorable combination. For the majority

TABLE 10.3

Estimated Size and Power^a for Each Combination of Randomization Rule and Analytic Approach, Based on 5000 Simulated Trials Derived from 5 SWOG Studies

Study & Allocation Rule ^b	Analysis Approach					
	Logrank		PH Model ^c		Stratified Logrank	
	Size	Power	Size	Power	Size	Power
<i>SWOG-8516</i>						
PB	0.022	0.834	0.023	0.843	0.022	0.812
PS	0.024	0.835	0.027	0.850	0.026	0.816
SB	0.020	0.847	0.023	0.859	0.024	0.834
<i>SWOG-8501</i>						
PB	0.025	0.836	0.026	0.841	0.026	0.821
PS	0.023	0.844	0.023	0.852	0.024	0.827
SB	0.027	0.831	0.027	0.832	0.027	0.819
<i>SWOG-9210</i>						
PB	0.023	0.788	0.024	0.791	0.024	0.773
PS	0.021	0.793	0.021	0.799	0.022	0.780
SB	0.023	0.803	0.024	0.809	0.024	0.794
<i>SWOG-9308</i>						
PB	0.019	0.836	0.024	0.884	0.022	0.871
PS	0.019	0.853	0.025	0.886	0.022	0.876
SB	0.023	0.843	0.028	0.886	0.027	0.875
<i>SWOG-8738</i>						
PB	0.023	0.842	0.026	0.851	0.021	0.871
PS	0.019	0.843	0.025	0.860	0.026	0.878
SB	0.023	0.848	0.024	0.860	0.026	0.877
Estimated type I error significantly less than the nominal 0.025						
Estimated power significantly lower than the best performing approach						

^a2 × SE (estimated size) = 0.004, 2 × SE (estimated power) = 0.011.

^bPB: permuted block; PS: Pocock–Simon adaptive allocation; SB: stratified permuted block.

^cPH: proportional hazards.

of the scenarios, the PH model regression performed slightly better than either the logrank or the stratified logrank, though the differences were generally quite small. In some settings, particularly *SWOG-8516* and *SWOG-8501*, there is a suggestion that the stratified logrank test does not provide as much power as the PH model based test, possibly related to the highly stratified nature of these trials.

In earlier work²⁴ we examined the effect of the number of covariates, their distribution in the study population, and the size of covariate effects on treatment effect inference. These simulations indicated that the number of covariates and their distribution have a smaller effect on the performance of these test statistics than the factors examined here. The magnitude of covariate effects was the key factor in the choice of analytic approach — the larger the effect, the more important it was to include that factor in the analysis and to model it correctly.

10.7 DISCUSSION

The motivation behind this study was to determine general recommendations for using covariates in the analysis of clinical trials employing randomization strategies that may or may not incorporate these factors. The approaches and scenarios examined via simulation were based on practical examples in which these very decisions are made on a regular basis. Though not comprehensive in examining the many factors that may have influence, these simulations provide a useful description of the magnitude of impact of the key aspects related to covariates in the design and analysis. In the settings examined, all combinations of randomization schemes and test statistics produced valid tests in the sense that the estimated type I error rates were never significantly greater than the nominal level. The type I error rates in analyses that ignored highly predictive covariates that had been used in the randomization scheme were considerably reduced, however, when highly predictive covariates were used in the randomization but not in the analysis.

These simulation studies indicate that in realistic settings, the use of covariates in treatment allocation may slightly improve the power. Based simply on these performance characteristics, we found little evidence to suggest the superiority of one type of constrained randomization rule over another. With this knowledge, the choice of randomization can be based primarily on other considerations including logistics and the desire to eliminate the remote possibility of extreme imbalance.

This work indicates that statistical power may be improved by including any strongly predictive covariates in the analysis, regardless of their use in the randomization. Further, there can be a substantial loss in power when highly predictive covariates were not accounted for in the analysis, especially if they were used in the randomization. Where PH assumptions for the covariates are reasonable, a PH model based test is expected to provide better power than a stratified analysis. In most settings examined, however, there was little to distinguish the two approaches. Even when PH assumptions were known to be violated, PH model based tests were valid and nearly as powerful as the more technically correct stratified logrank tests.

In summary, use of covariates in the randomization scheme or analyses did not affect the validity of the treatment arm comparisons, regardless of the analytic approach. The impact of their use on the power of a study was generally modest. Only when highly predictive covariates were used in the randomization but not in the analysis did we find evidence of marked degradation in performance (conservative type I error rates and a significant loss of power). In both the hypothetical trials and those derived from completed SWOG studies, we found that stratified block and Pocock–Simon adaptive randomizations schemes paired with either stratified logrank tests or tests from a PH regression with adjustment for covariates provided similar operating characteristics for the test.

ACKNOWLEDGMENT

This work was supported in part by grant ROI-CA90998 from the National Cancer Institute.

REFERENCES

1. Box, G.E.P., Hunter, W.G., and Hunter, J.S. *Statistics for Experimenters*. Wiley, New York, 1978.
2. Kalish, L.A. and Begg, C.B. Treatment allocation methods in clinical trials: a review. *Stat. Med.*, 4, 129, 1985.
3. Hill, A.B. The clinical trial. *Br. Med. Bull.*, 7, 278, 1951.
4. Pocock, S.J. and Simon, R. Sequential treatment assignment with balancing for prognostic factors in controlled clinical trials. *Biometrics*, 31, 103, 1975.
5. Begg, D.B. and Iglewicz, B. A treatment allocation procedure for sequential clinical trials. *Biometrics*, 36, 81, 1980.
6. Taves, D.R. Minimization: a new method of assigning patients to treatment and control groups. *Clin. Pharmacol. Ther.*, 15, 443, 1974.
7. Gail, M.H., Wieand, S., and Piantadosi, S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71, 431, 1984.
8. Lagakos, S.W. and Schoenfeld, D.A. Properties of proportional-hazards score tests under misspecified regression models. *Biometrics*, 40, 1037, 1984.
9. Struthers, C.A. and Kalbfleisch, J.D. Misspecified proportional hazards models. *Biometrika*, 78, 363, 1986.
10. Anderson, G.L. and Fleming, T.R. Model misspecification in proportional hazards regression. *Biometrika*, 82, 527, 1995.
11. Meinert, C.L. *Clinical Trials: Design, Conduct and Analysis*. Oxford University Press, New York, 1986, 193.
12. Friedman, L.M., Furberg, C.D., and DeMets, D.L. *Fundamentals of Clinical Trials*. John Wright PSG Inc., Boston, 1981, 165.
13. Peto, R., et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient (I). *Br. J. Cancer*, 34, 585, 1976.
14. Green, S.B. and Byar, D.B. The effect of stratified randomization on size and power of statistical tests in clinical trials. *J. Chron. Dis.*, 31, 445, 1978.
15. Rubin, D. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Stat.*, 2, 1, 1977.
16. Rubin, D. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.*, 6, 34, 1978.
17. Cox, D.R. Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. B*, 34, 187, 1972.
18. Kalbfleisch, J.D. and Prentice, R.L. *The Statistical Analysis of Failure Time Data*. 2nd ed. Wiley, New York, 2002, 107.
19. Fisher, R.I., et al. Comparison of a standard regimen (CHOP) with three intensive chemotherapy regimens for advanced non-Hodgkin's lymphoma. *N. Engl. J. Med.*, 328, 1002, 1993.
20. Alberts, D.S., et al. Intraperitoneal cisplatin plus intravenous cyclophosphamide versus intravenous cisplatin plus intravenous cyclophosphamide for stage III ovarian cancer. *N. Engl. J. Med.*, 335(26), 1950, 1996.
21. Berenson, J.R., et al. Maintenance therapy with alternate-day prednisone improves survival in multiple myeloma patients. *Blood*, 99, 3163, 2001.
22. Wozniak, A.J., et al. Randomized trial comparing cisplatin with cisplatin plus vinorelbine in the treatment of advanced non-small-cell lung cancer: a Southwest Oncology Group study. *J. Clin. Oncol.*, 16, 2459, 1998.

23. Gandara, D.R., et al. Evaluation of cisplatin intensity in metastatic non-small-cell lung cancer: a phase III study of the Southwest Oncology Group. *J. Clin. Oncol.*, 11, 873, 1993.
24. Anderson, G.L. Mismodelling covariates in Cox regression. Ph.D. thesis. University of Washington, Seattle, WA, 1989.

11 Factorial Designs with Time to Event Endpoints

Stephanie Green, Ph.D.

CONTENTS

11.1	Introduction	181
11.1.1	Approach 1	182
11.2.2	Approach 2	183
11.1.3	Approach 3	184
11.2	Simulation Study	184
11.3	Example: Southwest Oncology Group Study 8300	186
11.4	Other Approaches to Multiarm Studies	187
11.5	Conclusion	188
	References	189

11.1 INTRODUCTION

The frequent use of the standard two-arm randomized clinical trial is due in part to its relative simplicity of design and interpretation. Conclusions are straightforward: either the two arms are shown to be different or they are not. Complexities arise with more than two arms; with four arms there are 6 possible pairwise comparisons, 19 ways of pooling and comparing two groups, 24 ways of ordering the arms, plus the global test of equality of all four arms. Some subset of these comparisons must be identified as of interest; each comparison has power, level, and magnitude considerations; the problems of multiple testing must be addressed; and conclusions can be difficult, particularly if the comparisons specified to be of interest turn out to be the wrong ones.

Factorial designs are sometimes considered when two or more treatments, each of which has two or more dose levels, possibly including level 0 or no treatment, are of interest alone or in combination. A factorial design assigns patients equally to each possible combination of levels of each treatment. If treatment i , $i = 1, \dots, K$, has l_i levels, the result is an $l_1 \times l_2 \times \dots \times l_K$ factorial. Generally, the aim is to study the effect of levels of each treatment separately by pooling across all other treatments. The assumption often is made that each treatment has the same effect regardless of assignment to the other treatments. Statistically, an assumption of no interaction is made.

The use of factorial designs in clinical trials has become common, as noted in McAlister et al.¹ Statistical papers on the topic include a theoretical discussion of

factorials in the context of the proportional hazards model presented by Slud.² Other recent contributions to the topic include those by Byar,³ who suggested potential benefit in use of factorials for studies with low event rates such as screening studies; Simon and Freedman,⁴ who discussed Bayesian design and analysis of 2×2 factorials, allowing for some uncertainty in the assumption of no interaction; Hung,⁵ who discussed testing first for interaction when outcomes are normally distributed and interactions occur only if there are effects of both treatment arms; another by Hung⁶ concerning testing for unbalanced factorial clinical trials; and by Akritas and Brunner,⁷ who proposed a non-parametric approach to analysis of factorial designs with censored data making no assumptions about interaction. On the applied side, Green et al.⁸ discussed limitations of factorial designs and McAlister et. al.¹ discussed the quality of analysis and reporting in recently published factorial trials.

The multiple comparisons problem is one of the issues that must be considered in factorial designs. If tests of each treatment are performed at level α , which is typical for factorial designs,⁹ then the experiment-wide level, defined as the probability that at least one comparison will be significant under the null hypothesis, is greater than α . There is disagreement on the issue of whether all primary questions should each be tested at level α or whether the experiment-wide level across all primary questions should be level α , but clearly if the probability of at least one false positive result is high, a single positive result from the experiment will be difficult to interpret and may well be dismissed by many as inconclusive. Starting with global testing followed by pairwise tests only if the global test is significant is a common approach to limit the probability of false positive results. A Bonferroni approach, where each of T primary tests is performed at α/T , is also an option.

Power issues also must be considered. From the point of view of individual tests, power calculations are straightforward under the assumption of no interaction — calculate power according to the number of patients in the combined groups. A concern even in this ideal case may be the joint power. For instance, in a 2×2 trial of observation (O) versus treatment A versus treatment B versus the combination AB, if power to detect a specified effect of A is $1-\beta$ and power to detect a specified effect of B is also $1-\beta$, the joint power to detect the effects of both is closer to $1-2\beta$.

From the point of view of choosing the best arm, power considerations are considerably more complicated. The best arm must be specified for the possible true configurations; the procedures for designating the preferred arm at the end of the trial, which generally is the point of a clinical trial, must be specified; and the probabilities of choosing the best arm under alternatives of interest must be calculated. Several approaches can be considered in the context of a 2×2 trial.

11.1.1 APPROACH 1

The first approach is to perform the analysis assuming there are no interactions, using two one-sided tests, A versus not-A and B versus not-B. For example, test $\alpha = 0$ and $\beta = 0$ in a two-parameter proportional hazards model $\lambda = \lambda_0 \exp(\alpha z_A + \beta z_B)$, where λ is the survival hazard rate and z_A and z_B are treatment indicators. If neither test is significant, O is assumed to be the preferred arm. If A is better than not-A and B versus not-B is not significant, then A is assumed to be the preferred arm. B is assumed

best if the reverse is true. If A is better than not-A and B is better than not-B, then AB is preferred.

11.2.2 APPROACH 2

The second approach is to first perform a two-sided test for interaction using the model $\lambda = \lambda_0 \exp(\alpha z_A + \beta z_B + \gamma z_A z_B)$. If the interaction term γ is not significant, then base conclusions on the tests of A versus not-A and B versus not-B as in Approach 1. If it is significant, then base conclusions on tests of the three terms in the model and on appropriate subset tests. The treatment of choice is as follows:

Arm O if

1. γ is not significant, A versus not-A is not significant, and B versus not-B is not significant, or
2. γ is significant and negative (favorable interaction), α and β are not significant in the three-parameter model, and the test of O versus AB is not significant, or
3. γ is significant and positive (unfavorable interaction) and α and β are not significant in the three-parameter model

Arm AB if

1. γ is not significant, and A versus not-A and B versus not-B are both significant, or
2. γ is significant and favorable and α and β are both significant in the three-parameter model, or
3. γ is significant and favorable, α is significant and β is not significant in the three-parameter model, and the test of A versus AB is significant, or
4. γ is significant and favorable, β is significant and α is not significant in the three-parameter model, and the test of B versus AB is significant, or
5. γ is significant and favorable, α is not significant and β is not significant in the three-parameter model, and the test of O versus AB is significant

Arm A if

1. γ is not significant, B versus not-B is not significant, and A versus not-A is significant, or
2. γ is significant and favorable, α is significant and β is not significant in the three-parameter model, and the test of A versus AB is not significant, or
3. γ is significant and unfavorable, α is significant and β is not significant in the three-parameter model, or
4. γ is significant and unfavorable, α and β are significant in the three-parameter model, and the test of A versus B is significant in favor of A

Arm B if

results are similar to A above but with the results for A and B reversed.

Arm A or Arm B if

γ is significant and unfavorable, α and β are significant in the 3 parameter model, and the test of A versus B is not significant.

11.1.3 APPROACH 3

The third approach is to control the overall level of the experiment by first doing an overall test of differences among the four arms, for example, the four-arm logrank test. Proceed with Approach 2 above only if this test is significant. If the overall test is not significant, then arm O is concluded to be the treatment of choice.

11.2 SIMULATION STUDY

To illustrate the issues in factorial designs, a simulation of a 2×2 factorial trial will be used. The simulated trial had 125 patients per arm accrued over 3 years with 3 additional years of follow-up. Survival was exponentially distributed on each arm, and median survival was 1.5 years on the control arm. The sample size was sufficient for a one-sided 0.05 level test of A versus no-A to have power 0.9 with no effect of B, no interaction, and an A/O hazard ratio of 1/1.33. Various cases were considered using the model $\lambda = \lambda_0 \exp(\alpha z_A + \beta z_B + \gamma z_A z_B)$: neither A nor B effective (α and $\beta = 0$), A effective ($\alpha = -\ln(1.33)$) with no effect of B, A effective and B detrimental ($\beta = \ln(1.5)$), and both A and B effective (α and β both $= -\ln(1.33)$). Each of these were considered with no interaction ($\gamma = 0$), favorable interaction (AB hazard improved compared to expected, $\gamma = -\ln(1.33)$), and unfavorable interaction (worse, $\gamma = \ln(1.33)$). For each case, 2500 realizations were used for estimating characteristics of the three approaches outlined above. Table 11.1 summarizes the cases considered.

The possible outcomes of a trial of O versus A versus B versus AB are to recommend one of O, A, B, AB, or A or B but not AB. Tables 11.2–11.4 show the simulated probabilities of making each of the conclusions in the twelve cases of Table 11.1, for the approach of ignoring interaction, for the approach of testing for interaction, and for the approach of doing a global test before testing for interaction. The global test was

TABLE 11.1
Median Survival Times from Arms $\begin{pmatrix} O & A \\ B & AB \end{pmatrix}$ Used in the Simulation. Each Case has the Median of the Best Arm in Bold

Interaction	Case							
	1: Null		2: Effect of A No Effect of B		3: Effect of A Effect of B		4: Effect of A Detrimental Effect of B	
No	1.5	1.5	1.5	2	1.5	2	1.5	2
interaction	1.5	1.5	1.5	2	2	2.67	1	1.33
Unfavorable	1.5	1.5	1.5	2	1.5	2	1.5	2
interaction	1.5	1.13	1.5	1.5	2	2	1	1
Favorable	1.5	1.5	1.5	2	1.5	2	1.5	2
interaction	1.5	2	1.5	2.67	2	3.55	1	1.77

TABLE 11.2

Simulated Probability of Conclusion with Approach 1: No Test of Interaction. Each Case has the Median of the Best Arm in Bold

Case, Interaction	Conclusion				
	O	A	B	AB	A or B
1, none	0.890	0.055	0.053	0.002	0
1, unfavorable	0.999	0.001	0	0	0
1, favorable	0.311	0.243	0.259	0.187	0
2, none	0.078	0.867	0.007	0.049	0
2, unfavorable	0.562	0.437	0	0.001	0
2, favorable	0.002	0.627	0.002	0.369	0
3, none	0.010	0.104	0.095	0.791	0
3, unfavorable	0.316	0.244	0.231	0.208	0
3, favorable	0	0.009	0.006	0.985	0
4, none	0.078	0.922	0	0	0
4, unfavorable	0.578	0.422	0	0	0
4, favorable	0.002	0.998	0	0	0

TABLE 11.3

Probability of Conclusion with Approach 2: Test of Interaction. Each Case has the Median of the Best Arm in Bold

Case, Interaction	Conclusion					Test for Interaction, Probability of Rejection
	O	A	B	AB	A or B	
1, none	0.865	0.060	0.062	0.005	0.008	0.116
1, unfavorable	0.914	0.036	0.033	0	0.017	0.467
1, favorable	0.309	0.128	0.138	0.424	0	0.463
2, none	0.086	0.810	0.006	0.078	0.019	0.114
2, unfavorable	0.349	0.601	0.001	0.001	0.048	0.446
2, favorable	0.003	0.384	0.002	0.612	0	0.426
3, none	0.009	0.089	0.089	0.752	0.061	0.122
3, unfavorable	0.185	0.172	0.167	0.123	0.353	0.434
3, favorable	0	0.004	0.003	0.990	0.002	0.418
4, none	0.117	0.883	0	0	0	0.110
4, unfavorable	0.341	0.659	0	0	0	0.472
4, favorable	0.198	0.756	0	0.046	0	0.441

done at the two-sided 0.05 level. Other tests were done at the one-sided 0.05 level, with the exception of tests of A versus B and $\gamma = 0$, which were done at the two-sided 0.1 level. For each table the probability of drawing the correct conclusion is in bold.

Tables 11.2–11.4 illustrate several points. In the best case of using Approach 1 when in fact there is no interaction, the experiment level is 0.11, and the power when

TABLE 11.4
Probability of Conclusion with Approach 3: Global Test Followed by Approach 2. Each Case has the Median of the Best Arm in Bold

Case, Interaction	Conclusion					Global Test, Probability of Rejection
	O	A	B	AB	A or B	
1, none	0.972	0.011	0.010	0.003	0.004	0.052
1, unfavorable	0.926	0.032	0.026	0	0.015	0.578
1, favorable	0.503	0.049	0.057	0.390	0	0.558
2, none	0.329	0.578	0.001	0.074	0.018	0.684
2, unfavorable	0.528	0.432	0	0	0.039	0.541
2, favorable	0.014	0.374	0.001	0.611	0	0.987
3, none	0.068	0.069	0.063	0.741	0.059	0.932
3, unfavorable	0.466	0.067	0.072	0.109	0.286	0.535
3, favorable	0	0.004	0.003	0.990	0.002	1.00
4, none	0.117	0.882	0	0	0	0.997
4, unfavorable	0.341	0.659	0	0	0	1.00
4, favorable	0.198	0.756	0	0.046	0	0.999

both A and B are effective is 0.79, about as anticipated and possibly insufficiently conservative. Apart from that, Approach 1 is best if there is no interaction. The probability of choosing the correct arm is reduced if Approach 2 testing first for interaction is used instead of Approach 1 in all four cases with no interaction.

If there is an interaction, Approach 2 may or may not be superior. If the interaction masks the effectiveness of the best regimen, it is better to test for interaction. See, for example, case 4 with an unfavorable interaction, where the difference between A and not-A is diminished due to the interaction. If the interaction enhances the effectiveness of the best arm, testing is detrimental; see case 4 with favorable interaction, where the difference between A and not-A is larger due to the interaction while B is still clearly ineffective. In all cases the power for detecting interactions is poor. Even using 0.1 level tests, the interactions were detected at most 47% of the time in these simulations.

Approach 3 does restrict the overall level, but this is at the expense of a reduced probability of choosing the correct arm when the four arms are not sufficiently different for the overall test to have high power.

Unfavorable interactions are particularly detrimental to a study. The probability of identifying the correct regimen is poor for all methods if the correct arm is not the control arm. Approach 1, assuming there is no interaction, is particularly poor.

11.3 EXAMPLE: SOUTHWEST ONCOLOGY GROUP STUDY 8300

As an illustration of how use of a factorial design may compromise a study, consider Southwest Oncology Group Study 8300, which is similar to case 4 with an unfavorable

interaction, reported by Miller et al.¹⁰ In this study in limited non-small cell lung cancer, the roles of both chemotherapy and prophylactic radiation to the brain were of interest. All patients received radiation to the chest and were randomized to receive prophylactic brain irradiation (PBI) versus chemotherapy versus both PBI and chemotherapy versus no additional treatment. PBI was to be tested by combining across the chemotherapy arms and chemotherapy was to be tested by combining across PBI arms. Investigators chose a Bonferroni approach to limit type I error. The trial design specified level 0.025 for two tests, a test of whether PBI was superior to no PBI and a test of whether chemotherapy was superior to no chemotherapy. No other tests were specified. It was assumed that PBI and chemotherapy would not affect each other. Unfortunately, PBI was found to be detrimental to patient survival. Although the interaction term was not significant, the worst arm was PBI plus chemotherapy, followed by PBI, then no additional treatment, then chemotherapy alone. Using the design criteria, one would conclude that neither PBI nor chemotherapy should be used. With this outcome, however, it was clear that the comparison of no further treatment versus chemotherapy was critical, but the study had seriously inadequate power for this test, and no definitive conclusion could be made concerning chemotherapy.

Once you admit your $K \times J$ factorial study is not one K arm study and one J arm study that happen to be in the same patients but rather a $K \times J$ arm study with small numbers of patients per arm, the difficulties become more evident. This is not changing the rules of the design but rather acknowledging the reality of most clinical settings. Perhaps in studies where A and B have unrelated mechanisms of action and are being used to affect different outcomes, assumptions of no biologic interaction may not be unreasonable. However, in general A cannot be assumed to behave the same way in the presence of B as in the absence of B. Potential drug interactions, overlapping toxicities, differences in compliance and other confounding factors all make it more reasonable to assume there will be differences. Furthermore, interaction may occur simply as an artifact of the particular model chosen. In the simulation, *no interaction* meant the O/A and B/AB hazard ratios were equal. If instead, *equally effective* is defined as equal absolute increase in median, then two of the no interaction cases in Table 12.1 turn into interaction cases. There is no biologic reason that compels any particular mathematical formulation of *equally effective* to be correct. Thus, lack of biologic rationale does not necessarily provide reassurance with respect to interaction according to the model identified for analysis. Models are rarely completely correct, so statistical interactions of modest size are likely even if there is no evidence of biologic interactions.

11.4 OTHER APPROACHES TO MULTIARM STUDIES

Various approaches to multi-arm studies are available. If the example study could be formulated as O versus A, B, and AB, as might be the case if lower doses of A and B are used for the combination, the problem of comparing control versus multiple experimental arms would apply. There is a long history of papers on this problem^{10–13} focusing on appropriate global tests or tests for sub-hypotheses. Liu and Dahlberg¹⁴ discussed design and provided sample size estimates based on the least favorable alternative for the global test for K -arm trials with time-to-event endpoints. Their procedure, a K -sample logrank test is performed at level α followed by α level pair-wise

tests if the global test is significant, has good power for detecting the difference between a control arm and the best treatment. These authors emphasized the problems when power is considered in the broader sense of drawing the correct conclusions. Properties are good for this approach when each experimental arm is similar either to the control arm or the best arm, but not when survival times are more evenly spread out among the control arm and other arms.

Designs for ordered alternatives are another possibility. For example, suppose there are theoretical reasons to hypothesize superiority of A over B resulting in the alternative $O < B < A < AB$. Liu et al.¹⁵ proposed a modified logrank test for

ordered alternatives, $T = \sum_{i=1}^{K-1} L_{(i)} / \left[\sum_{i=1}^{K-1} L_{(i)} + 2 \sum_{i < j} \text{cov}(L_{(i)}, L_{(j)}) \right]^{1/2}$ where $L_{(i)}$ is

the numerator of the one-sided logrank test between the pooled groups $1, \dots, i$ and pooled groups $i + 1 \dots K$. This test is used as the global test before pairwise comparisons. Similar comments apply as to the more general case above, with the additional problem that the test will not work well if the ordering is mis-specified. A related approach includes a preference ordering, say by expense of the regimens, which at least has a good chance of being specified correctly, and application of a bubble sort analysis. For example, the most costly treatment is preferred only if significantly better than the rest, the second most costly only if significantly better than the less costly arms and if the most costly is not significantly better, and so on. This approach is discussed by Chen and Simon.¹⁶

Any model assumption that is incorrect can result in problems. As with testing for interactions, testing any assumptions can either be beneficial or detrimental, with no way of ascertaining beforehand which is the case. If assumptions are tested, procedures must be specified for when the assumptions are shown not to be met, which changes the properties of the experiment and complicates sample size considerations. Southwest Oncology Group study S8738¹⁷ provides an example of incorrect assumptions. This trial randomized patients to low-dose CDDP versus high-dose CDDP versus high dose CDDP plus Mitomycin-C, with the obvious hypothesized ordering. The trial was closed approximately halfway through the planned accrual because survival on high dose CDDP was convincingly shown *not* to be superior to standard dose CDDP by the hypothesized 25%. In fact, it appeared to be worse. A beneficial effect of adding Mitomycin-C to high dose CDDP could not be ruled out at the time, but this comparison became meaningless in view of the standard-dose versus high-dose comparison.

11.5 CONCLUSION

The motivation for simplifying assumptions in multi-arm trials is clear. The sample size required to have adequate power for multiple plausible alternatives, while at the same time limiting the overall level of the experiment, is large. If power for specific pairwise comparisons is important for any outcome, then the required sample size is larger. An even larger sample size is needed if detection of interaction is of interest. To detect an interaction of the same magnitude as the main effects in a 2×2 trial, four times the sample size is required,¹⁸ thereby eliminating what most view as the primary advantage to factorial designs.

Likely not all simplifying assumptions are wrong, but disappointing experience tells us the risk is not negligible. Unfortunately, the small sample sizes resulting from over-simplification may lead to unacceptable chances of inconclusive results. The correct balance between conservative assumptions versus possible efficiencies is rarely clear.

In the case of factorial designs, combining treatment arms seems to be a neat trick, yielding multiple answers for the price of one, until one starts to consider how to protect against the possibility that the assumptions underlying the trick are incorrect.

REFERENCES

1. McAlister, F., Straus, S., Sackett, D., and Altman D. Analysis and reporting of factorial trials: a systematic review. *J. Am. Medical Assoc.*, 289, 2545–2553, 2003.
2. Slud, E. Analysis of factorial survival experiments. *Biometrics*, 50, 25–38, 1994.
3. Byar, J. Factorial and reciprocal control design. *Stat. Med.*, 9, 55–64, 1990.
4. Simon, R. and Freedman L. Bayesian design and analysis of two $\times$ two factorial clinical trials. *Biometrics*, 53, 456–464, 1997.
5. Hung, H. Two-stage tests for studying monotherapy and combination therapy in two-by-two factorial trials. *Stat. Med.*, 12, 645–660, 1993.
6. Hung, H. Evaluation of a combination drug with multiple doses in unbalanced factorial design clinical trials. *Stat. Med.*, 19, 2079–2087, 2000.
7. Akritas, M. and Brunner, E. Nonparametric methods for factorial designs with censored data. *J. Am. Stat. Assoc.*, 92, 568–576, 1997.
8. Green, S., Liu, P.Y., and O'Sullivan, J. Factorial design considerations. *J. Clin. Oncology*, 20, 3424–3430, 2002.
9. Gail M., et al. Factorial trial of three interventions to reduce the progression of pre-cancerous gastric lesions in Sandong, China: design issues and initial data. *Controlled Clin. Trials*, 19, 352–369, 1998.
10. Miller, T. et al. A randomized trial of chemotherapy and radiotherapy for stage III non-small cell lung cancer. *Cancer Therapeutics*, 1, 229–236, 1998.
11. Dunnet, C. A multiple comparisons procedure for comparing several treatments with a control. *J. Am. Stat. Assoc.*, 60, 573–583, 1955.
12. Marcus, R., Peritz, E., and Gabriel K. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63, 655–660, 1976.
13. Tang, D.-I. and Lin, S. An approximate likelihood ratio test for comparing several treatments to a control. *J. Am. Stat. Assoc.*, 92, 1155–1162, 1997.
14. Liu, P.Y. and Dahlberg, S. Design and analysis of multiarm clinical trials with survival endpoints. *Controlled Clin. Trials*, 16, 119–130, 1995.
15. Liu, P.Y., Tsai, W.Y., and Wolf M. Design and analysis for survival data under order restrictions with a modified logrank test. *Stat. Med.*, 17, 1469–1479, 1998.
16. Chen, T. and Simon, R. Extension of one-sided test to multiple treatment trials. *Controlled Clin. Trials*, 15, 124–134, 1994.
17. Gandara, D. et al. Evaluation of cisplatin in metastatic non-small cell lung cancer: a Phase III study of the Southwest Oncology Group. *J. Clin. Oncology*, 11, 873–878, 1993.
18. Peterson, B. and George, S. Sample size requirements and length of study for testing interaction in a $2 \times K$ factorial design when time to failure is the outcome. *Controlled Clin. Trials*, 14, 511–522, 1993.

12 Noninferiority Trials

*Kenneth J. Kopecky and
Stephanie Green*

CONTENTS

12.1	Introduction	191
12.2	Hypothesis Tests and Confidence Intervals	192
12.3	Sample Size Determination	193
12.3.1	Noninferiority Trial with Binary Outcome	193
12.3.2	Noninferiority Trial with Time-to-Event Outcome	195
12.4	Example	197
12.5	Determining the Noninferiority Margin	198
12.6	Benefit of E Compared to No Treatment.....	199
12.7	Example (Continued).....	202
12.8	Intention-to-Treat vs. Per-Protocol	203
12.9	Discussion	203
	References	204

12.1 INTRODUCTION

Phase III therapeutic trials in oncology are conducted to compare the effectiveness of treatment regimens. In most settings an accepted standard therapy exists, and the motivation is the hope that a new treatment regimen E will prove to be superior to the standard therapy S in some respect, for example survival or response rate. Such trials are called superiority trials. Let $\rho(X, Y)$ denote a parameter that characterizes the difference between outcomes with treatments X and Y . Without loss of generality, we assume that $\rho(E, S)$ is parameterized so that $\rho(E, S) = 0$ if E and S are equally effective and $\rho(E, S) > 0$ if E is superior to S . For example, for comparisons of response rates, $\rho(E, S)$ might be rate difference $P_E - P_S$ or the log odds ratio $\ln\{P_E(1 - P_S)/[P_S(1 - P_E)]\}$, where P_X denotes the response rate for arm $X = S$ or E . Similarly $\rho(E, S)$ might be the log hazard ratio for S relative to E in comparing survival. In the hypothesis testing context, a superiority trial tests the null hypothesis $H_0: \rho(E, S) \leq 0$ against the one-sided alternative $H_A: \rho(E, S) > 0$.

In some circumstances, however, a trial is motivated by the hope that E is nearly as effective as S in some respect. In cancer trials such studies can be of particular interest because many treatment regimens have significant detrimental consequences, for example severe and even life-threatening toxicity, high cost, inconvenience, or difficulty of administration that reduces adherence to the regimen. A new

regimen that is less toxic, expensive, or difficult but nearly as effective as the standard regimen may therefore be the preferred treatment approach. Trials in this setting can be classified as either equivalence trials or noninferiority trials.

The term *equivalence trial* has generally come to refer to trials for which the hypothesis of interest is that results with E are similar to S , i.e., neither better nor worse beyond reasonable limits. In general such trials are of little value for studies of the effectiveness of new treatments because there would be little reason to reject E if it proved significantly more effective than S . Equivalence trials will not be discussed in this paper.

Noninferiority trials, in contrast, are inherently one-sided. The hypothesis of interest is that E is, at worst, inferior to S by a defined margin. In the hypothesis testing context, the null hypothesis for a noninferiority trial is therefore $H_0: \rho(E, S) \leq M$ against the one-sided alternative $H_A: \rho(E, S) > M$ where $M < 0$ is the allowable margin of inferiority, i.e., the maximum loss of effectiveness that is considered acceptable. M , referred to herein as the noninferiority margin, is sometimes called the equivalence limit.¹

The last several years have seen a great deal of research into the methods for designing, analyzing, and interpreting noninferiority trials; see for example the January 30, 2003, issue of *Statistics in Medicine* (volume 22, number 2). In this paper we address some practical issues regarding noninferiority trials in the setting of cancer therapy research.

12.2 HYPOTHESIS TESTS AND CONFIDENCE INTERVALS

There are two general approaches to analyzing the results of noninferiority trials: significance testing and confidence interval (CI) estimation. The distinction between the two approaches is to some extent artificial. Although these are sometimes viewed as alternative approaches to statistical analysis, they provide complementary information and are indeed often used together. Nevertheless the design and analysis of noninferiority trials is generally presented in terms of one or the other approach.

In the hypothesis testing approach, the aim is to perform a significance test of the null hypothesis of unacceptable inferiority, $H_0: \rho(E, S) \leq M$ (recall $M < 0$ is the largest allowable margin of inferiority), against the one-sided alternative $H_A: \rho(E, S) > M$. If the null hypothesis is rejected, then noninferiority is concluded. The study's statistical power is the probability of correctly rejecting the null hypothesis of unacceptable inferiority. Ordinarily the study's sample size is determined to ensure that this test has a high level of power, typically 90% or 80%, if E and S are indeed equally effective, i.e., under the specific alternative hypothesis $H_A: \rho(E, S) = 0$.

The aim of the confidence interval approach is to calculate a lower confidence limit $\hat{\rho}^L(E, S)$ for $\rho(E, S)$. In this paper, we assume $\hat{\rho}^L(E, S)$ is the lower limit of a two-sided $100(1 - \alpha)\%$ confidence interval. If $M < \hat{\rho}^L(E, S)$, then the study's results are inconsistent with unacceptable inferiority, and noninferiority within the margin M is concluded. If, however, $\hat{\rho}^L(E, S) \leq M$, then the possibility of unacceptable inferiority cannot be ruled out. The sample size is determined to ensure a high level of power, in other words, a high probability that $\hat{\rho}^L(E, S)$ will exceed M if in fact E is not inferior to S . As in the hypothesis testing approach, it is usually reasonable to calculate sample size under the assumption that $\rho(E, S) = 0$, i.e., that E and S are equally effective.

As is generally the case, the two approaches are closely related. For example, one can define significance tests on the basis of confidence intervals including or excluding hypothesized values. Nevertheless the CI approach provides more information than the simple result of the hypothesis test because it gives a range of plausible values for the treatment difference parameter $\rho(E, S)$. Therefore the development below is based on the CI approach.

In the following sections, sample size formulas are presented for simple noninferiority trials in which the treatments are compared with respect to a binary outcome, such as response to therapy, or to a time-to-event outcome, such as overall or disease-free survival. These relatively simple models may be appropriate for noninferiority trials of cancer therapies in many if not most settings. Indeed models of such simplicity have been used to design a number of published cancer trials, although as described in the examples below, they have not always been used with sufficient attention to the selection of appropriate design specifications.

12.3 SAMPLE SIZE DETERMINATION

12.3.1 NONINFERIORITY TRIAL WITH BINARY OUTCOME

For a binary outcome variable such as complete response, a noninferiority trial can be based on the difference in outcome probabilities: $\rho(E, S) = P_E - P_S$. The $100(1 - \alpha)\%$ two-sided confidence limits for $P_E - P_S$ are given by

$$\hat{P}_E - \hat{P}_S \pm z_{\alpha/2} \sqrt{\frac{\hat{P}_E(1-\hat{P}_E)}{N_E} + \frac{\hat{P}_S(1-\hat{P}_S)}{N_S}}$$

where N_x and $\hat{P}_x$ are the number of patients and proportion of patients with the outcome, respectively, in arm $x = E$ or S , and z_p denotes the 100 p th percentile of the standard normal distribution. For a noninferiority trial comparing E and S , one wants a high probability (power) that the lower confidence limit will be greater than the noninferiority margin M if E is indeed not inferior. The sample size required to ensure statistical power of $100(1 - \beta)\%$ if the true event probabilities are P_E and P_S may therefore be calculated using the formula

$$N = \left[\frac{z_{\alpha/2} + z_\beta}{M + (P_E - P_S)} \right]^2 \times \left[\frac{P_E(1-P_E)}{K_E} + \frac{P_S(1-P_S)}{1-K_E} \right] \quad (12.1)$$

where N is the total number of patients and K_E is the proportion randomized to E . Ordinarily trials are designed to have adequate power to reject inferiority if the regimens are equally effective, that is, if $P_E = P_S = P$ for a value of P specified by the investigators. Also for any fixed P , M , α , and β , N is minimized when $K_E = 0.5$. With these assumptions Equation (12.1) simplifies to

$$N = 4 \times \left[\frac{z_{\alpha/2} + z_\beta}{M} \right]^2 \times [P(1-P)] \quad (12.2)$$

Table 12.1 displays the numbers of patients required for various combinations of P , M , α , and β , calculated using Equation (12.2).

Alternatively, the difference in event rates can be expressed as the log odds ratio: $\rho(E, S) = \ln\{P_E(1 - P_S)/[P_S(1 - P_E)]\}$.² The estimated log odds ratio is

$$\hat{\rho}(E, S) = \ln\left(\frac{\hat{P}_E}{1 - \hat{P}_E}\right) - \ln\left(\frac{\hat{P}_S}{1 - \hat{P}_S}\right)$$

which has estimated standard error

$$\sqrt{\frac{1}{N_E \hat{P}_E (1 - \hat{P}_E)} + \frac{1}{N_S \hat{P}_S (1 - \hat{P}_S)}}$$

For noninferiority margin M , now expressed as the log odds ratio, the sample size formula analogous to Equation (12.1) is

$$N = \left[\frac{z_{\alpha/2} + z_{\beta}}{M - \ln(\text{OR})} \right]^2 \times \left[\frac{1}{K_E P_E (1 - P_E)} + \frac{1}{(1 - K_E) P_S (1 - P_S)} \right] \tag{12.3}$$

where $\ln(\text{OR}) = \ln[P_E/(1 - P_E)] - \ln[P_S/(1 - P_S)]$ is defined by the values of P_E and P_S under the alternative hypothesis. If this is chosen as the hypothesis of equivalence ($P_E = P_S = P$) and $K_E = 0.5$, then Equation (12.3) simplifies to

$$N = 4 \times \left(\frac{z_{\alpha/2} + z_{\beta}}{M} \right)^2 \times \left[\frac{1}{P(1 - P)} \right] \tag{12.4}$$

Table 12.2 displays the numbers of patients required for selected values of P , M , α , and β , calculated using Equation (12.4). If $P_S > 0.5$, and therefore $P > 0.5$, then for any fixed values of M , α , and β , the sample size increases with increasing P . This occurs because the absolute difference between P_E and P_S corresponding to the log odds ratio M decreases.

TABLE 12.1
Study Sizes Required for Noninferiority Comparison of Binary Outcome Based on Arithmetic Difference in Outcome Rates

Probability of Event with Standard Therapy	Noninferiority Margin M , Expressed as Additive Change in Event Rate	Confidence Level *			
		95%		90%	
		Statistical Power			
		90%	80%	90%	80%
0.50	−0.05	4204	3140	3426	2474
	−0.10	1052	786	858	620
0.75	−0.05	3154	2356	2570	1856
	−0.10	790	590	644	464
0.90	−0.05	1514	1132	1234	892
	−0.10	380	284	310	224

* Confidence level for two-sided confidence interval.

TABLE 12.2
Study Sizes Required for Noninferiority Comparison of Binary Outcome Based on Odds Ratio

Probability of Event with Standard Therapy (P_S)	Noninferiority Margin, Expressed as Odds Ratio [$M = \text{Log}$ odds ratio] (P_E^{**})	Confidence Level *			
		95%		90%	
		Statistical Power			
		90%	80%	90%	80%
0.50	0.80 [−0.22] (0.44)	3377	2523	2752	1987
	0.70 [−0.36] (0.41)	1322	988	1078	778
0.75	0.80 [−0.22] (0.71)	4502	3363	3670	2649
	0.70 [−0.36] (0.68)	1763	1317	1437	1037
0.90	0.80 [−0.22] (0.88)	9379	7006	7644	5519
	0.70 [−0.36] (0.86)	3671	2743	2992	2160

* Confidence level for two-sided confidence interval.
** Event probability for arm E assuming probability P_S for arm S and the indicated $E:S$ odds ratio.

12.3.2 NONINFERIORITY TRIAL WITH TIME-TO-EVENT OUTCOME

For cancer clinical trials in which the outcome of interest is time until some event such as death or disease progression, it is often reasonable for sample size calculations to assume that times to events are exponentially distributed. Assume N patients will accrue at a uniform rate from time 0 until A , and that follow-up will continue for an additional period of duration F , at which point observation will be censored for all patients remaining alive (censoring at time $A + F$). Assuming E and S are exponentially distributed, the hazard rates can be parameterized as $\lambda_E = e^\alpha$ and $\lambda_S = e^{\alpha + \rho(E, S)}$, respectively, so that the log hazard ratio for S relative to E , $\rho(E, S)$, is greater than 0 if E is superior to S , and the noninferiority hypothesis is $H_0: \rho(E, S) \leq M$ for a specified $M < 0$. For example, in Southwest Oncology Group study SWOG-8412, which compared carboplatin plus cyclophosphamide to the standard therapy of cisplatin plus cyclophosphamide with respect to overall survival of patients with stage III or IV ovarian cancer, the margin was chosen to correspond to a mortality hazard ratio (E relative to S) of 1.3, corresponding to a noninferiority margin $M = -\ln(1.3) = -0.262$.³

Letting $\rho = \rho(E, S)$, the maximum likelihood estimator of the log hazard ratio is

$$\hat{\rho} = \ln(d_E/T_E) - \ln(d_S/T_S)$$

with estimated standard error

$$SE(\hat{\rho}) = \sqrt{(d_E + d_S)/d_E d_S}$$

where d_x and T_x are the number of events and total observation time (sum of all times from study entry until observation of the event or censoring), respectively, observed on arm $x = E$ or S . Thus the $100(1 - \alpha)\%$ confidence limits are given by

$$\hat{\rho} \pm z_{\alpha/2} \times SE(\hat{\rho})$$

The total number of patients required to ensure at least $100(1 - \beta)\%$ probability that the lower limit of the two-sided $100(1 - \alpha)\%$ confidence interval will exceed M when the true value of the log hazard ratio is $\rho(E, S)$ can be calculated as

$$N = \left(\frac{z_{\alpha/2} + z_{\beta}}{M - \rho(E, S)} \right)^2 \times \left[\frac{1}{K_E Q_E} + \frac{1}{(1 - K_E) Q_S} \right] \tag{12.5}$$

where K_E is the proportion of patients randomized to E and Q_E and Q_S are the expected proportions of patients whose deaths will be observed in the two arms. For exponentially distributed times to events

$$Q_x = 1 - \frac{\exp(-\lambda_x F) \times [1 - \exp(-\lambda_x A)]}{\lambda_x A} \tag{12.6}$$

for $x = E$ or S . Typically the study requires adequate power under the alternative hypothesis $\rho(E, S) = 0$, in which case $\lambda_E = \lambda_S$, N is again minimized if $K_E = 0.5$, and Equation (12.5) simplifies to

$$N = \left(\frac{4}{Q_S} \right) \left(\frac{z_{\alpha/2} + z_{\beta}}{M} \right)^2 \tag{12.7}$$

The value of λ_S , and of λ_E if needed, can be determined from parameters of the anticipated true exponential distributions; for example, if the median event time is predicted to be $T_{0.5}$, then $\lambda_x = \ln(2)/T_{0.5}$. Note from Equation (12.6) that the required number of patients depends on the accrual and followup times through their ratios to the median event time, $A/T_{0.5}$ and $F/T_{0.5}$. A JavaScript program that calculates N using Equation (12.5) for the alternative hypothesis that the true value of $\rho(E, S)$ is 0 and for arbitrary Q_E and Q_S is available at <http://www.swogstat.org/stat/public/equivsurv.htm>.

Table 12.3 displays examples of the sample size required for various combinations of M , α , and β , calculated using Equations (12.6) and (12.7) assuming $A/T_{0.5} = 4$ and $F/T_{0.5} = 1$. Whether an accrual goal can be met within an accrual period depends of course on the accrual rate. For example, if the noninferiority margin is set at 1.10 and the study uses the two-sided 95% confidence level and requires 90% power if the true hazard ratio is 1.0, then the study must accrue 5569 patients in

TABLE 12.3
Study Sizes Required for Noninferiority Comparison of Exponentially Distributed Time-to-Event Outcome Based on Hazard Ratio

Noninferiority Margin, Expressed as E:S Hazard Ratio (corresponding M)	Duration of study periods (expressed as ratios to median event time with S)		Confidence Level *			
			95%		90%	
	Actual	Followup	Statistical Power			
			90%	80%	90%	80%
1.05 (−0.049)	4	1	21249	15873	17319	12503
1.10 (−0.095)	4	1	5569	4160	4539	3277
1.30 (−0.262)	4	1	735	549	599	433

* Confidence level for two-sided confidence interval.

a period that is four times the median survival time with standard therapy. If the median is 1.5 years, then an accrual rate of about 929 per year is required for 4 years. Having specified the noninferiority margin, confidence level, and power, a combination of A , F , and accrual rate, if any is indeed feasible, can be found by trial and error.

In general, increasing the duration of follow-up will not reduce the study size greatly. For example with margin 1.10, confidence level 95%, and power 90% for true hazard ratio 1.0, increasing the follow-up from 1 to 4 times the median survival with standard therapy decreases the required number of patients to from 5569 to 4727. This is because the power depends heavily on NQ_S , the expected number of patients whose events will be observed. For the specifications in Table 12.3, Q_S increases from 0.83 if $F = 1$ to 0.98 if $F = 4$, a modest increase. Note that increasing F further above 4 would clearly yield little additional power. Shortening the accrual interval increases the necessary number of patients and, consequently, the required accrual rate. For example, under the same specifications as above, total accrual of 6343 would be required for the study to complete accrual in a period twice, rather than four times, as long as the median event time. Note that this more than doubles the required accrual rate.

The results in Table 12.3 show that the sample sizes required to demonstrate noninferiority within reasonable margins may be very large. For outcomes such as death or, for highly lethal diseases, disease progression, it may be difficult to justify a 30% or even 10% increase in mortality as an acceptable trade-off for some other benefit such as decreased toxicity or cost. Moreover allowing a large noninferiority margin increases the risk that an apparently noninferior E will not be significantly superior to no treatment. However, it may not be possible to conduct an adequately powered noninferiority trial targeting a smaller margin. The likelihood that the accrual period will need to extend for several multiples of the median time to event may render a noninferiority trial impractical. This is especially true for studies of patients with good prognosis for long median event times, a circumstance in which noninferiority questions may be most likely to arise. Shorter accrual periods may be appropriate for studies in patients with poor prognosis; however, in such settings, superiority trials aimed at improving prognosis are likely to have much higher research priority than noninferiority trials.

12.4 EXAMPLE

Two recent papers provide interesting examples of noninferiority trials.^{4,5} Each study compared the experimental treatment, zolendronic acid ($E = \text{Zol}$), to the standard therapy, pamidronate ($S = \text{Pam}$), with respect to skeletal-related events (SREs, i.e., skeletal metastases or osteolytic lesions) in cancer patients or hypercalcemia related to malignancy (HCM). The primary endpoints for the noninferiority comparisons were binary outcomes: avoidance of SRE (excluding HCM) during treatment and follow-up in the first trial (the SRE trial), and complete response (CR) of HCM in the second trial (the HCM trial). Since these two trials did not address the same endpoint, they cannot be compared directly. Nevertheless they present an interesting contrast.

Both studies were designed to test noninferiority of Zol by comparing the lower limit of a two-sided $100(1 - \alpha)\%$ confidence interval for the difference in event rates

$(P_{Zol} - P_{Pam})$ to a specified margin M . For the SRE trial, the lower limit of the 90% confidence interval for $P_{Zol} - P_{Pam}$ was required to exceed -0.08 in order to conclude noninferiority of *Zol*, i.e., $\alpha = 0.1$ and $M = -0.08$. The trial was designed to ensure 80% power ($\beta = 0.2$) if the true probabilities were $P_{Zol} = P_{Pam} = 0.5$ and therefore from Equation (12.2) required $N = 968$ patients. The HCM trial was designed with very different specifications that led to a much smaller study size: the criterion for noninferiority of *Zol* required that the lower limit of the 95% confidence interval exceed a margin of $M = -0.10$, and the trial was designed to ensure 80% power if the true event probabilities were $P_{Zol} = 0.92$ and $P_{Pam} = 0.90$. Consequently from Equation (12.1) the HCM trial had a goal of only $N = 180$ patients.

Ordinarily noninferiority trials are designed to have adequate power to reject inferiority if the two regimens are equivalent. However the HCM trial was designed to have adequate power if *Zol* was in fact slightly superior to *Pam*. Had the power been targeted to the conventional alternative hypothesis that *Zol* is equivalent to *Pam*, i.e., $P_{Zol} = P_{Pam} = 0.90$, the study would have required $N = 284$ patients (Table 12.1) rather than 180. Therefore, the HCM trial was seriously underpowered and had an unacceptably high probability of failing to reject the hypothesis of *Zol*'s inferiority if it was in fact equivalent to *Pam*.

12.5 DETERMINING THE NONINFERIORITY MARGIN

Determining an appropriate value for the noninferiority margin M requires careful clinical and statistical consideration. Wiens described general approaches for selecting noninferiority margins.¹ One is to choose a margin small enough to ensure that if E is nearly as effective as S , then it is reasonable to conclude that E is superior to no treatment or placebo (P). To do this, one might try to choose a specific value of M that is close enough to 0 to provide assurance of E 's benefit; this appears to be the approach taken in the two *Zol* vs. *Pam* trials described above. Alternatively, one may choose to define M as a fraction of the benefit of S compared to P : $M = -\phi\rho(S, P)$ where $0 < \phi < 1$. This way of defining M has an appealingly intuitive interpretation: the noninferiority margin corresponds to preserving $100(1 - \phi)\%$ of the benefit of S relative to P . The general problem of assessing benefits of S and/or E relative to P is discussed further below. In either case, this may be a useful approach when there is very strong evidence that S is very effective. However, if the statistical significance or the magnitude of the benefit of S is modest, then a noninferiority margin chosen solely by this approach may not provide the desired assurance that E has some benefit.

A second approach described by Wiens is to choose a margin based solely on clinical importance, that is, a difference that is judged to be clinically unimportant. Choosing a margin in this way is clearly fraught with subjectivity and uncertainty. Nevertheless it may be a reasonable approach when there is little information available from which to predict the likely benefit of S compared to P in the planned trial.

For example, Michallet et al. designed a noninferiority comparison of $E =$ pegylated interferon alpha-2b (rIFN- α 2b) compared to $S =$ rIFN- α 2b as therapy for chronic myelogenous leukemia (CML) in chronic phase.⁶ They specified a 20% decrease in the odds of major cytogenetic response (MCR) as the noninferiority margin. Although the authors did not describe their rationale for this specification, it was

perhaps justified by a combination of the two approaches above. In particular, regarding Weins' first approach, it is plausible to assume that spontaneous MCR is a very unlikely event, so if S is clearly beneficial relative to P and the trial showed that E preserved a large proportion of the benefit of S , then it might be reasonable to conclude that E is sufficiently beneficial to replace S . However, the trial was unlikely to reach such a conclusion for two reasons. First, the rate of MCR with S was not very high: 20% in the investigators' assumption for design purposes. Thus even if E could be shown to preserve a large proportion of the benefit of S , it probably would remain unclear whether E had sufficient benefit compared to no treatment to warrant its use. Moreover, the study as designed was badly underpowered. The sample size was calculated to ensure that the lower limit of a two-sided 95% confidence interval for the log odds ratio had 80% probability of exceeding a noninferiority margin corresponding to an odds ratio of 0.8 if the true MCR rates were 20% with S and 30% with E ; that is, under an alternative hypothesis that E was substantially superior to S . Using Equation (12.3) or a similar formula, the study was designed to include 300 patients, although 344 were actually included. The trial was therefore badly underpowered: a total of 3941 patients would be required to have 80% power if E and S were in fact equivalent with MCR rates of 20%. As it turned out, the study failed to reject the hypothesis of inferiority. However, due to the study's low power, this result is in fact inconclusive.

The third approach described by Wiens is to select a margin that ensures the distributions of outcomes for patients on the E and S arms are similar in some respect other than the parameter for which they are being compared.¹ A simple example of this is the use of the difference in rates to determine the noninferiority margin for the odds ratio in trials with binary outcome variables: the difference in rates may be more readily understood by clinicians than the odds ratio. For example, in the CML trial described above, the noninferiority margin corresponded to a 20% decrease in the odds of MCR, and the MCR rate was assumed to be 20% with S .⁶ The noninferiority margin therefore corresponds to a decrease in the MCR rate from 20% with S to 17% with E , which is easy to interpret clinically.

12.6 BENEFIT OF E COMPARED TO NO TREATMENT

In order for E to be a suitable replacement for S on the basis of noninferiority, it is important to consider not only (A) whether E is nearly as effective as S , but also (B) whether E is superior to no treatment or placebo (P), that is, whether $\rho(E, P) > 0$. The ideal equivalence trial would randomize to all three approaches, P , S , and E , provided it is ethical to randomize to P . This might be the case if evidence supporting S is weak, such as the lower bound of confidence interval being close to 0, uncertainty due to poorly conducted previous trials, changes over time, short term effects, or highly variable differences in treatment effect across previous studies. Koch and Röhmel describe conduct of gold-standard noninferiority trials.⁷ The experimental regimen E is accepted if E is significantly better than P and if it is noninferior to S . S itself need not be significantly better than P . Hypotheses are ordered hierarchically. $H_{01}: \rho(E, P) = 0$ vs. $\rho(E, P) > 0$ is tested in step one and, if rejected, $H_{02}: \rho(E, S) = M$ vs. $\rho(E, S) > M$ is tested in step two, each typically at the 0.025 level. If both hypotheses are rejected, the

trial is concluded successful. If the trial is successful, further tests can be done to address questions about the superiority of S compared to P and of E compared to S without compromising family-wise type-one error.

If a treatment of established effectiveness exists (i.e., S), it would be unethical to randomize patients to the nontreatment arm P . In this circumstance only historical experience may be available to address consideration (B). Therefore the benefit of E relative to P must be inferred from the current trial comparing E to S and from whatever information is available regarding the benefit of S relative to P . The latter inference typically relies on the assumption of some kind of constancy condition, i.e., that the benefit of S compared to P observed in prior studies carries over to the current noninferiority trial.^{8,9} If the noninferiority margin is defined as a fraction of the benefit of S compared to P , $M = -\phi\rho(S, P)$ for a specified value of ϕ , then the need to infer $\rho(S, P)$ from historical experience also affects the inference regarding consideration (A).

Some approaches to estimating or testing $\rho(E, P)$ are described briefly below. First, however, it is important to recognize that the validity of inferences concerning the benefit of E relative to P in a trial with no P arm can be very sensitive to the validity of the constancy condition.^{10–12} Therefore, the constancy condition requires careful, critical consideration. Any number of effects may operate to violate the constancy condition: patterns of referral of patients to study centers may change; diagnostic criteria may change; supportive care may become more effective. Moreover variations in the design and conduct of the current and prior trials may invalidate the constancy condition: differences in eligibility criteria, response criteria, adherence to treatment regimens and follow-up requirements, subsequent “rescue” therapy for treatment failures, and many other factors may reduce the actual effect of S compared to P .⁹ These are basically the same pitfalls that arise in any comparison of current to historical data. The potential for error in this extrapolation is of particular concern in studies of cancer therapies because there may be few or even no historical studies that can provide a solid basis for estimating the benefit of S compared to no treatment.

To reduce the bias that can arise from imputing the standard regimen’s effectiveness using historical placebo-controlled studies, the current study should be as similar as possible to the prior successful trials.⁹ However, a high degree of similarity may be difficult to achieve. For example in adult AML (excluding AML-M3), the most widely used remission induction regimens, including ara-C and an anthracycline such as daunorubicin or idarubicin, have been arguably standard therapy for two decades or more. During that time there have been no placebo-controlled trials of AML remission induction chemotherapy. Moreover it has been proposed recently to revise the diagnostic criteria for AML to include a condition, RAEB-T, that was previously classified as one of the myelodysplastic syndromes. This reclassification was based on evidence that RAEB-T and AML patients under age 60 have similar prognoses for overall survival.¹³ Such a revision makes it even more difficult for future noninferiority trials of chemotherapy for AML, if any are attempted, to infer reliably the benefits of S and E relative to P .

Simon has argued that noninferiority trials cannot produce reliable results unless there is very strong evidence that the benefit of S compared to P is large.¹⁴ This would not be the case if, for example, the benefit of S was established in

placebo-controlled trials with marginal levels of statistical significance or, equivalently, confidence intervals for the magnitude of the benefit barely excluded zero. Similarly, suppose S is the latest in a series of two or more standard therapies that have been adopted sequentially based on trials without placebo controls that have each demonstrated marginally significant incremental superiority over the previous standard regimen. In such situations it may be extremely difficult to infer the benefit of S relative to no treatment. In particular if the current S was adopted on the basis of a noninferiority trial, or perhaps a series of noninferiority trials, then its benefit relative to P may be diminished, a phenomenon that has been called biocreep.⁸ D'Agostino et al. suggest dealing with biocreep by always using the “best” regimen as the active control arm; however, this may not be practical: selection of the best arm may need to rely on nonrandomized historical comparisons, and a new trial in which standard care is a therapy that is no longer widely used may fail to accrue patients. Moreover, biocreep in the opposite direction might also occur in a series of superiority trials as a result of publication bias: if superiority trials with statistically significant results are more likely to be published than those without significant results, then the benefit of S may be overestimated. This kind of biocreep would be less likely if positive results of superiority trials were routinely investigated in confirmatory trials.

Several approaches have been proposed for what is sometimes called putative placebo analysis, which is used for inferring the comparison of E to P from the current noninferiority trial of E versus S and historical data regarding the benefit of S relative to P .

Fisher¹⁵ and Hasselblad and Kong¹⁶ exploited the fact that if treatment differences can be represented on an additive scale, then

$$\rho(E,P) = \rho(E,S) + \rho(S,P) \quad (12.8)$$

Note that this may require transformation to an additive scale, for example to log odds ratios or log hazard ratios. If the two terms on the right side of Equation (13.8) are estimated from independent studies, we also have

$$\text{var}[\hat{\rho}(E,P)] = \text{var}[\hat{\rho}(E,S)] + \text{var}[\hat{\rho}(S,P)] \quad (12.9)$$

The terms on the right sides of Equations (12.8) and (12.9) can be estimated from the current noninferiority trial $[\rho(E,S)]$ and the historical data $[\rho(S,P)]$. For example $\rho(S,P)$ and $\text{var}[\hat{\rho}(S,P)]$ might be estimated from a meta-analysis of prior trials of S vs. P . Thus $\rho(E,P)$ and its variance can be estimated, and the hypothesis $H_0: \rho(E,P) \leq 0$ can be tested against the one-sided alternative $H_A: \rho(E,P) > 0$, or a confidence interval for $\rho(E,P)$ can be calculated to support inference regarding the benefit of E compared to P .

Rothman described a “two confidence interval” approach in which a lower confidence limit for the effect of E relative to S , $\hat{\rho}^L(E,S)$, is compared to a multiple of an upper confidence limit for the benefit of S relative to P , $\delta \hat{\rho}^U(S,P)$, based on historical data. The confidence level of the latter interval must be chosen to ensure the desired probability of type I error.¹²

Simon described a Bayesian approach to analysis after a trial of E vs. S in which the expected responses for P , S , and E are μ , $\mu + \eta$, and $\mu + \theta$,

respectively.^{14, 17} In terms of Equation, (12.8) $\theta = \rho(E, P)$ and $\eta = \rho(S, P)$. Thus, assuming positive values of η and θ indicate benefit, E would be of interest if $\eta > 0$ (S has benefit) and $\theta > (1 - \varphi)\eta$ for a specified value of φ between 0 and 1 (E preserves at least $100(1 - \varphi)\%$ of S 's benefit). Note that the latter condition corresponds to $\theta - \eta > -\varphi\eta$, and the noninferiority margin for $\rho(E, S) = \theta - \eta$ is therefore $M = -\varphi\eta$, which is proportional to $\eta = \rho(S, P)$. Through straightforward Bayesian analysis, the posterior joint distribution of (μ, η, θ) is estimated, allowing in turn estimation of the posterior probabilities of the events $\{\eta > 0\}$ and $\{\theta > (1 - \varphi)\eta\}$, and perhaps more usefully of the event $\{\eta > 0 \text{ and } \theta > (1 - \varphi)\eta\}$. Simon also describes a variation of this model for the proportional hazards regression model assuming the hazard ratios for S and E relative to P are $\exp(\eta)$ and $\exp(\theta)$, respectively.^{14, 17}

Note that in Simon's approach there is no attempt to compare E to P directly as in the approach of Hasselblad and Kong. Instead the benefit of E compared to P is inferred from the findings that (A) S is superior to P , and (B) E retains at least $100(1 - \varphi)\%$ of the benefit of S . If the results do not favor finding (A), that is, if the posterior probability of $\{\eta > 0\}$ is small, then inference regarding finding (B) is of little interest and the validity of the constancy condition is in doubt; however, the posterior probability of $\{\theta > 0\}$ may still provide useful information about the benefit of E . Allowing for uncertainty in μ and η is appealing, but as always with Bayesian approaches, care must be taken that choice of priors does not unduly influence the conclusion concerning E .

12.7 EXAMPLE (CONTINUED)

For neither the SRE trial nor the HCM trial did the study's reported design or implementation provide for testing or estimating the benefit of *Zol* compared to a putative untreated control. No rationale for this omission was provided. The investigators may have assumed that the benefits of *Pam* are sufficiently large and precisely estimated that showing *Zol* is almost as effective as *Pam* would ensure that it must have benefit compared to no treatment. The risk in this assumption was clearly shown by the results of the HCM trial, in which the CR rate with *Pam*, 69.7%, was markedly lower than had been expected based on earlier HCM trials.⁵ Had the trial been designed to have 80% power if $P_{Zol} = P_{Pam} = 0.7$, it would have required $N = 660$ patients rather than the 180 required by targeting the power to $P_{Zol} = 0.92$ and $P_{Pam} = 0.90$.

In further analysis of the SRE results, an estimate of *Pam*'s benefit was derived from three prior placebo-controlled studies.¹⁸ It was estimated that *Pam* increased the rate of SRE avoidance from the placebo's 48.0% to 61.1%, an improvement of 13.1% with 95% confidence interval (7.3%, 18.9%). In the SRE trial, the estimated event rates were 56% with *Zol* and 54% with *Pam*, for a difference of 2% with 95% confidence interval (-3.7%, 7.9%). Comparing these confidence intervals, it was concluded that *Zol* retained at least $(7.3 - 3.7)/7.3 = 49.3\%$ of *Pam*'s benefit relative to placebo. Had the lower confidence limit for *Zol*'s inferiority been less than -7.3%, the study could not have ruled out the possibility that *Zol* had no benefit compared to placebo. Again it was noteworthy that *Pam* was less effective in the SRE trial (54%) than might have been predicted from the earlier placebo-controlled trials (61.1%), raising doubt about the validity of the constancy condition.

As it turned out, the proportions of patients with favorable outcomes were higher with *Zol* than with *Pam* in both of the SRE and HCM trials. While this may obviate some of the concerns about these trials' limitations as noninferiority studies, it must be emphasized that these trials were at risk of producing inconclusive or even misleading results concerning their stated noninferiority objectives.

12.8 INTENTION-TO-TREAT VS. PER-PROTOCOL

The principle of making treatment comparisons on the basis of intention to treat (ITT) is widely accepted for superiority trials because the inclusion of ineligible or untreated patients, lack of adherence to treatment regimens, and other inevitable flaws in study conduct are generally expected to increase the noise in the study and may reduce the apparent benefit, if any, of the experimental treatment. That is, comparisons of treatment groups based on ITT are expected to be conservative in superiority trials. In contrast, *per-protocol* (PP) analysis is limited to eligible patients who received treatment according to protocol. PP analysis, which may be unbiased with regard to treatment differences, may have severely limited generalizability. For noninferiority trials, however, the situation regarding ITT analysis may be reversed: by increasing the noise and reducing the apparent difference between *E* and *S*, the flaws in study conduct may bias the study toward a conclusion of noninferiority.^{9,19} Thus ITT may be anti-conservative for noninferiority trials. Jones et al. recommend carrying out both ITT and PP analyses and careful examination of the patients who are excluded from PP analysis in order to investigate the impact on the anticonservatism of ITT analysis.⁹

12.9 DISCUSSION

It is widely understood that that failure to reject the null hypothesis that two regimens have equivalent effectiveness does not constitute proof that they are equivalent. However, this understanding is not universal. One still hears, for example, estimated survival curves that are quite close together (described sometimes as superimposable) interpreted as evidence of equivalence, with no mention of the variability inherent in the estimates. In a similar vein, a large *p*-value for a test of the hypothesis of equality may also be taken inappropriately to imply equivalence; for example, $p = 0.9$ may be misinterpreted to mean that there is a 90% chance the null hypothesis is true. Krysan and Kemper reviewed 25 randomized controlled trials that claimed equivalence of case mortality between treatments for bacterial meningitis in children and found that the claim was based on absence of significant superiority in 23 trials, and that only three of the 25 trials had adequate power to rule out a 10% difference in mortality.²⁰ While noninferiority questions may frequently arise in cancer therapy research, the questions of whether and how such questions should be addressed require careful consideration.

The conduct, analysis, and interpretation of noninferiority trials have been the subject of extensive methodological research in recent years, and a number of valuable insights and methods have resulted. However, in the context of cancer therapy trials, relatively simple methods may be adequate and even preferred. Much of the recent

methodological research has addressed the problem of inferring whether E is superior to P , and the validity of this inference is heavily dependent on the validity of a constancy condition that permits extrapolation from prior studies comparing S to P . In the setting of cancer therapy trials, it may be very difficult to justify this extrapolation. In that setting, if a noninferiority trial is necessary and appropriate, a practical approach may be the following: define a fixed noninferiority margin M that preserves a sufficient fraction of the benefit of S relative to P , based on the best judgment of the magnitude of that benefit and the best clinical opinion as to the fraction of that benefit that must be preserved; then perform an adequately powered trial to produce a hypothesis test or confidence interval to support an inference about whether E is or is not inferior to S .

The requirement that noninferiority trials be adequately powered would seem to be self evident. And yet, as the examples above illustrate, even published studies have failed to meet this minimal quality requirement. Adequate power may require extraordinarily large study sizes, especially for studies of highly significant outcomes such as survival or, for some diseases, response to therapy. These sample sizes can be reduced in either of two ways. First the margin of noninferiority, M , can be chosen to have a comparatively large magnitude. However, such a value of M may represent an unacceptable trade-off for the benefits that the experimental treatment offers. The other way to reduce study size is to require that the study have adequate power to reject inferiority only if E is in fact superior to S , as was done in the HCM and CML trials described above. However, this requires one to accept less than adequate power to conclude noninferiority if the two treatments are in fact equivalent. In other words, it has too large a probability of missing the benefits that accrue from using E in place of S if E and S are in fact equivalent.

Noninferiority trials requiring large numbers of patients may be difficult to complete; however, smaller trials having insufficient statistical power should be avoided because, as is true of any statistical study, they are too likely to produce inconclusive results. It remains true, as Simon pointed out in the previous edition of this handbook¹⁴ that superiority trials remain the preferred means to improve cancer therapy whenever possible.

REFERENCES

1. Wiens, B.L. Choosing an equivalence limit for noninferiority or equivalence trials. *Contr. Clin. Trials*, 23, 1, 2002.
2. Wang, H., Chow, S., and Li, G. On sample size calculation based on odds ratio in clinical trials. *J. Biopharm. Stat.*, 12, 471, 2002.
3. Alberts, D.S. et al. Improved therapeutic index of carboplatin plus cyclophosphamide versus cisplatin plus cyclophosphamide: final report by the Southwest Oncology Group of a phase III randomized trial in stages III and IV ovarian cancer. *J. Clin. Oncol.*, 10, 706, 1992.
4. Rosen, L.S. et al. Zoledronic acid versus pamidronate in the treatment of skeletal metastases in patients with breast cancer or osteolytic lesions of multiple myeloma: a phase III, double-blind, comparative trial. *Cancer J.*, 7, 377, 2001.
5. Major, P. et al. Zoledronic acid is superior to pamidronate in the treatment of hypercalcemia of malignancy: a pooled analysis of two randomized, controlled clinical trials. *J. Clin. Oncol.*, 19, 558, 2001.

6. Michallet, M. et al. Pegylated recombinant interferon alpha-2b vs recombinant interferon alpha-2b for the initial treatment of chronic-phase chronic myelogenous leukemia: a phase III study. *Leukemia*, 18, 309, 2004.
7. Koch, A., and Röhmle, J. Hypothesis testing in the “gold standard” design for proving the efficacy of an experimental treatment relative to placebo and a reference. *J. Biopharm. Stat.* 14, 315, 2004.
8. D’Agostino, R.B., Massaro, J.M., and Sullivan, L.M. Non-inferiority trials: concepts and issues - the encounters of academic consultants in statistics. *Stat. Med.*, 22, 169, 2003.
9. Jones, B. et al. Trials to assess equivalence: the importance of rigorous methods. *BMJ* 313, 36, 1996.
10. Wang, S.J., Hung, H.M.J., and Tsong, Y. Utility and pitfalls of some statistical methods in active controlled clinical trials. *Contr. Clin. Trials*, 23, 15, 2002.
11. Hung, H.M.J. et al. Some fundamental issues with noninferiority testing in active controlled trials. *Stat. Med.*, 22, 213, 2003.
12. Rothman, M. et al. Design and analysis of non-inferiority mortality trials in oncology. *Stat. Med.*, 22, 239, 2003.
13. Bennett, J.M. World Health Organization classification of the acute leukemias and myelodysplastic syndrome. *Int. J. Hemat.*, 72, 131, 2000.
14. Simon, R. Therapeutic equivalence trials, in *Handbook of Statistics in Clinical Oncology*, Crowley J., ed., Marcel Dekker, Inc., New York, 2001, chap. 10.
15. Fisher, L. Active control trials: what about a placebo? A method illustrated with clopidogrel, aspirin, and placebo. *J. Amer. Coll. Cardiol.*, 31, 49A, 1998.
16. Hasselblad, V. and Kong, D.F. Statistical methods for comparison of placebo in active-control trials. *Drug Inf. J.*, 35, 435, 2001.
17. Simon, R. Bayesian design and analysis of active control clinical trials. *Biometrics*, 55, 484, 1999.
18. Ibrahim, A. et al. Approval summary for zoledronic acid for treatment of multiple myeloma and cancer bone metastases. *Clin. Cancer Res.*, 9, 2394, 2003.
19. CDER/CBER. Guidance for Industry. E9 Statistical Principles for Clinical Trials, Food and Drug Administration, 1998.
20. Krysan, D.J. and Kemper A.R. Claims of equivalence in randomized controlled trials of the treatment of bacterial meningitis in children. *Pediatr. Infect. Dis. J.*, 21, 753, 2002.

13 Power and Sample Size for Phase III Clinical Trials of Survival

Jonathan J. Shuster, Ph.D.

CONTENTS

13.1	Clinical Summary	208
13.2	Introduction	209
13.3	A Basic Sample Size Formulation	210
13.3.1	Binomial Example (One-Sided test)	210
13.4	Exponential Survival	211
13.5	Applications to the Logrank Test and Cox Regression	212
13.6	Planning a Study with Piecewise Exponential Survival	213
13.6.1	Parameters to Identify	213
13.6.1.1	Input Parameters	213
13.6.1.2	Output Parameters	213
13.6.2	Key Observations	214
13.7	Sequential Monitoring by the O'Brien–Fleming Method	215
13.8	Evaluation of the Information Fraction	216
13.9	Multiple Treatments and Stratification	217
13.10	Competing Losses	218
13.11	Practical Conclusions	218
	Acknowledgment	219
	References	219
	Appendix I:	220
	Appendix II:	221
	Appendix III:	222

13.1 CLINICAL SUMMARY

Suppose you wish to conduct a clinical trial to compare the survival of leukemia patients under two regimens (one a standard or control therapy, and the other experimental). From past experience, suppose the 5-year survival under the standard therapy is 60%, and you would like your trial to be large enough to have a high probability of obtaining a statistically significant difference (i.e., high enough power) when the experimental therapy represents a 10% improvement to 70% survival at 5 years. Due to sampling error, even if the true target population rates are 60% and 70% at 5 years for the standard and experimental therapies respectively, the observed 5-year rates will be different. An undersized study may well miss this clinically important improvement. Conversely, an oversized study may waste resources and could delay both publication of the conclusions and the opening of a successor trial.

The benchmark for demonstrating efficacy is the randomized controlled trial where members of a prospective cohort are assigned randomly to one of the two treatments. Changes in concomitant care or population shifts could bias any comparison between two sequential series (historical controls), the first one treated on the standard therapy, the subsequent one on the experimental therapy.

This chapter contains the necessary information for biostatisticians to plan for randomized studies that allow for sequential monitoring of efficacy. This is a collaborative process between the clinical leadership and the biostatistician. Planning parameters needed are the following: (1) annual accrual rate; (2) minimum planned follow-up (e.g., 5 years) where the survival is pegged; (3, 4) planning survival at that time under the standard and experimental therapies; (5) sidedness of the test (A one-sided test is used, for example, if the experimental therapy is the control therapy plus a new component, and where the management decision to consider the new component as part of a future standard rests with the significance of the trial in favor of the experimental therapy; a two-sided test should be used when the better treatment is sought in a symmetric fashion); (6) the size of the type I error (α), the probability of falsely obtaining a significant difference (in the predicted direction if one-sided) when the two therapies are equivalent; (7) the size of the type II error (β), the probability of reaching an inconclusive result (non-significant) under the planning parameters for the study (for example a 10% improvement); (8) hazard factor, the average risk ratio of after to before the minimum planned follow-up time. (For example, if the risk of dying on a given day after 5 years on average is 60% of that of a day before 5 years, then the planning hazard factor is 0.6); and (9) loss to follow-up rate for the study.

Two technical points need to be mentioned. First, the logrank statistical test used in this chapter compares the entire survival curves, not just the fixed term survival (e.g., at 5 years). It asks if the true target survival curves for the experimental and standard treatments are the same. Second, the sample size calculation presumes proportional hazards. This means that if for patients alive at the start of day 103 the risk of failing on day 103 under the experimental therapy is 70% of that of the standard therapy, then the risk of failure on the experimental therapy is 70% of the standard therapy on every day. Both risks can vary from day to day, but the ratio is presumed

constant. Under the null hypothesis that the two treatments are equivalent, the risk ratio, experimental to standard, is of course 100% on every day, and hence proportional hazards hold. However, there are some situations where proportional hazards are not to be expected. For example, in studies of bone marrow transplantation vs. chemotherapy, there could well be more early deaths on the transplant arm with a later crossover in risk. Such studies could be planned for fixed term survival using Kaplan-Meier¹ actuarial estimates with planning based on a binomial (all or none response) model, per Section 13.3, adjusted for losses to follow-up per Section 13.10.

13.2 INTRODUCTION

This chapter is devoted to two treatment 1–1 randomized trials where the outcome measure is survival, or more generally, time until an adverse event. In practical terms, planning requires far more in the way of assumptions than trials whose results are accrued quickly. This chapter is concerned with the most common set of assumptions, proportional hazards, which implies that the ratio of the instantaneous probability of death at a given instant (treatment A: treatment B) for patients alive at time T is the same for all values of T . Sample size guidelines are presented first for situations in which there is no sequential monitoring. These are modified by an inflation factor to allow for O'Brien–Fleming type² sequential monitoring. The methodology initially is built upon a two-piece exponential model, but this is later relaxed to cover proportional hazards in general.

One of the most important aspects of this chapter centers on the nonrobustness of the power and type I error properties when proportional hazards are violated. This is especially problematic when sequential methods are employed. Statisticians are faced with the need for such methods on the one hand but with the reality that they are based on questionable forecasts on the other. As a partial solution to this problem, a leapfrog approach is proposed where a trial accrues a block of patients and is then put on hold.

In Section 13.3, a general sample size formulation for tests that are inversions of confidence intervals will be presented. In Section 13.4, results for exponential survival and estimation of the difference between hazard rates will be given. In Section 13.5, the connection between the exponential results and the logrank test as described in Peto and Peto³ and Cox regression⁴ are given. Section 13.6 is devoted to a generalization of the results for the exponential distribution to the two-piece exponential distribution with proportional hazards and to a numerical example. In Section 13.7, the necessary sample size is developed for the O'Brien–Fleming method of pure sequential monitoring. In that section, it is shown that the maximum sample size inflation over trials with no monitoring is typically of the order of 7% or less. For group sequential plans, this inflation factor is even less. Section 13.8 is devoted to ways of obtaining the information fraction time, an essential ingredient for sequential monitoring. The fraction of failures observed (interim analysis to final planned analysis) is the recommended measure. If that measure is indeed used and if the trial is run to a fixed number of failures, then the piecewise

exponential assumption can be relaxed further to proportional hazards. Section 13.9 deals with application of the methods to more complex designs, including multiple treatments and 2×2 factorial designs. Section 13.10 is concerned with competing losses. Finally, Section 13.11 gives the practical conclusions, including major cautions about the use of sequential monitoring.

13.3 A BASIC SAMPLE SIZE FORMULATION

The following formulation is peculiar looking but very useful. See Shuster⁵ for further details. Suppose the following equations (a) and (b) hold under the null (H_0) and alternative hypothesis (H_1) respectively:

$$\begin{aligned} H_0: \Delta &= \Delta_0 \text{ vs. } H_1: \Delta = \Delta_1 > \Delta_0 \\ \text{(a) } P[(\hat{\Delta} - \Delta_0)/SE > Z_\alpha] &\approx \alpha \\ \text{(b) } P[W + (\hat{\Delta} - \Delta_1)/SE > -Z_\beta] &\approx 1 - \beta \end{aligned}$$

where $W = (Z_\alpha + Z_\beta)(S - SE)/SE$. SE is a standard error estimate calculated from the data only, valid under both the null and alternate hypotheses; S is a function of the parameters (including sample size) under the alternate hypothesis, and S satisfies the implicit sample size equation:

$$S = (\Delta_1 - \Delta_0)/(Z_\alpha + Z_\beta) \quad (13.1)$$

Then it follows under H_1 ,

$$P[(\hat{\Delta} - \Delta_0)/SE > Z_\alpha] \approx 1 - \beta \quad (13.2)$$

That is, the approximate α level test has approximate power $1 - \beta$.

Note that Z_α and Z_β are usually (but need not always be) the 100α and 100β upper percentiles of the standard normal cumulative distribution function (CDF).

Suppose you have a confidence interval inversion test with normal distributions after standardizing by the standard error statistic. To validate the implicit sample size formula, you need show that under H_1 , S/SE converges to one in probability.

For two-sided tests, set up so that $\Delta_1 > \Delta_0$, α is replaced by $\alpha/2$.

13.3.1 BINOMIAL EXAMPLE (ONE-SIDED TEST)

For binomial trials with success rates P_1 and P_2 and equal sample size N per group,

$$\begin{aligned} S &= \text{sqrt}[\{P_1(1 - P_1) + P_2(1 - P_2)\}/N] \\ \Delta_1 &= (P_1 - P_2) \\ \Delta_0 &= 0 \end{aligned}$$

and hence from (13.1), *each* treatment will have sample size:

$$N = [P_1(1 - P_1) + P_2(1 - P_2)]\{(Z_\alpha + Z_\beta)/(P_1 - P_2)\}^2$$

This formula is useful for determining the sample size based on the Kaplan–Meier statistic.

For the two-sided test, we replace α by $\alpha/2$ in the above expression.

13.4 EXPONENTIAL SURVIVAL

If the underlying distribution is exponential, it can be shown that the stochastic process that plots the number of deaths observed (*Y*-axis) vs. total accumulated time on test (*X*-axis) is a homogeneous Poisson process with hazard rate equal to the constant hazard of the exponential distribution.

For treatments $i = 1, 2$ let

$$\lambda_i = \text{Hazard rate for treatment } i$$

$$F_i = \text{Total number of failures to total accumulated time on test } T_i$$

$$\hat{\lambda}_i = F_i/T_i \sim \text{Asy } N[\lambda_i, \lambda_i^2/E(F_i)]$$

$$\text{Let } \hat{\Delta} = \hat{\lambda}_1 - \hat{\lambda}_2 \text{ and } \Delta = \lambda_1 - \lambda_2.$$

In the notation of the previous section, one has

$$SE = [(\hat{\lambda}_1^2/F_1) + (\hat{\lambda}_2^2/F_2)]^{0.5}$$

and

$$S = [\lambda_1^2/E(F_1) + \lambda_2^2/E(F_2)]^{0.5} \quad (13.3)$$

If patients are accrued uniformly (Poisson arrival) over calendar time $(0, X)$ and followed until death or to the closure (time $X + Y$) with Y being the minimum follow-up time, then the probability of death for a random patient assigned to treatment i is easily obtained as:

$$P_i = 1 - Q_i \quad (13.4)$$

where

$$Q_i = \exp(-\lambda_i Y)[1 - \exp(-\lambda_i X)]/(\lambda_i X)$$

The expected number of failures on treatment i is

$$E(F_i) = 0.5\psi XP_i \quad (13.5)$$

where ψ is the accrual rate for the study (half assigned to each treatment). If the allocations are unequal with γ_i assigned to treatment i ($\gamma_1 + \gamma_2 = 1$), simply replace

the 0.5 in Equation (13.5) by γ_i . This is useful in planning studies of prognostic factors or clinical trials where the experimental treatment is very costly compared to the control.

After substituting Equation (13.3) in Equation (13.1), the resulting equation must be solved iteratively (bisection is the method of choice) to identify the accrual period X required for given planning values of the accrual rate ψ , minimum follow-up period Y , and values λ_1 and λ_2 (and hence of Δ) under the alternate hypothesis.

Similar methods for exponential survival have been published by Bernstein and Lagakos,⁶ George and Desu,⁷ Lachin,⁸ Morgan,⁹ Rubinstein et al.,¹⁰ and Schoenfeld.¹¹ These methods utilize various transformations and thus yield slightly different though locally equivalent results. The Schoenfeld paper allows the incorporation of covariates, while the Bernstein and Lagakos paper allows one to incorporate stratification.

13.5 APPLICATIONS TO THE LOGRANK TEST AND COX REGRESSION

Two important observations extend the utility of the above sample size formulation to many settings under proportional hazards.

1. Peto and Peto³ demonstrated the full asymptotic local efficiency of the logrank test when the underlying survival distributions are exponential. This implies that the power and sample size formulas of Section 13.4 apply directly to the logrank test (as well as to the likelihood based test used above.) See [Appendix I](#) for further discussion of the efficiency of the logrank test. For two treatments, the test for no treatment effect in Cox regression (with no other covariates) is equivalent to the logrank test.
2. Two distributions have proportional hazards if and only if there exists a continuous strictly monotonic increasing transformation of the time scale that simultaneously converts both to exponential distributions.

This means that an investigator can plan any trial that assumes proportional hazards as long as a planning transformation that converts the outcomes to exponential distributions is prespecified. This can be approximated well if historical control data are available. The only problems are to redefine the λ_i and to evaluate $E(F_i)$ taking the transformation into account because accrual may not be uniform in the transformed time scale.

Three papers offer software that can do the calculations: Halperin and Brown,¹² Cantor,¹³ and Henderson et al.¹⁴ These papers also investigate the robustness of the sample size to deviations from distributional assumptions. Another approach, that is quite robust to failure in correctly identifying the transformation presumes a two-piece exponential model (with two pieces). This is discussed in the next section along with a numerical example.

13.6 PLANNING A STUDY WITH PIECEWISE EXPONENTIAL SURVIVAL

13.6.1 PARAMETERS TO IDENTIFY

13.6.1.1 Input Parameters

ψ = annual planning accrual rate (total) (e.g., 210 patients per year are planned, 50% per arm)

Y = minimum follow-up time (e.g., $Y = 3$ years)

R_1 = Planning Y -year survival under the control treatment (e.g., $R_1 = 0.60$)

R_2 = Planning Y -year survival under the experimental treatment (e.g., $R_2 = 0.70$)
Sidedness of test (one or two) (e.g., one)

α = type I error (e.g., 0.05)

$\pi = 1 - \beta$, the power of the test. (e.g., 0.80)

ρ = posttime Y :pretime Y hazard ratio (e.g., 0.5)

ρ represents the piecewise exponential. Before the minimum follow-up Y , if the hazard is λ_i on treatment i , it is $\rho\lambda_i$ on treatment i after Y .

Note that $\lambda_i = -\ln(R_i)/Y$

13.6.1.2 Output Parameters

X = accrual period (e.g., 2.33 years)

$X + Y$ = total study duration (e.g., 5.33 years)

ψX = total accrual required (e.g., 490 patients)

$E(F_1) + E(F_2)$ = total expected failures (e.g., 196)

Based on observation 2 in Section 13.5, the calculation of $E(F_i)$ for Equations (13.3) and (13.5) is straightforward. For a randomly selected patient, the time at risk is uniformly distributed over the period from Y to $(X + Y)$, and the probability of death $D_i(x)$ conditional on a time at risk $Y + x$ is the probability of death by time Y , $(1 - R_i)$ plus the probability of surviving to time Y but dying between times Y and $x + Y$; that is,

$$D_i(x) = (1 - R_i) + R_i(1 - \exp(-\rho\lambda_i x))$$

The unconditional probability of death for a randomly selected patient on treatment i is found by taking the expectation of $D_i(x)$ over the uniform distribution for x from 0 to X .

For this piecewise exponential model and Poisson accrual process, the value of Q_i to be used in Equations (13.4) and (13.5) is:

$$Q_i = R_i[1 - \exp(-\rho\lambda_i X)]/(\rho\lambda_i X) \quad (13.6)$$

Note that λ_i is defined via the equation:

$$R_i = \exp(-\lambda_i Y) \quad (13.7)$$

and hence for exponential data ($\rho = 1$) the value of Q_i agrees with that below Equation (13.4).

13.6.2 KEY OBSERVATIONS

1. The expected failures on treatment i in the first Y years of patient risk is $0.5\psi X(1 - R_i)$ where ψ is the annual accrual rate, X is the accrual period, and R_i is the planning Y -year survival on treatment i .
2. If one transformed the subset of the time scale $(0, Y)$, $(0, Y)$ with a strictly monotonic increasing function, the expected number of failures that occurred before patient time Y would be unchanged. This in turn implies that the exponential assumption over the interval $(0, Y)$ can be relaxed to require proportional hazards only. (From Equation (13.7), the definition of $\lambda_i = -\ln(R_i)/Y$ in Equation (13.3) depends only on the fixed Y -year survival on treatment i , R_i , and is not affected by the transformation.)
3. All other things being equal, Equation (13.3) implies that the larger the expected number of failures, the greater the power of the test. Since the larger the value of ρ the greater the hazard post Y -years, power increases as ρ increases.
4. If hazard rates are low after time Y , the exponentiality assumption post Y is not important (although proportional hazards may be). It is convenient to think of ρ as an average hazard ratio (post:pre year Y) because it enters the sample size equation only through the expected number of failures. In fact, if one used the above sample size methodology to approximate the accrual duration and minimum follow-up but actually ran the trial until the total failures equaled the total expected in the study, $E(F_1) + E(F_2)$, then the power would hold up under proportional hazards without regard to the piecewise exponential assumption. To further illustrate this point, in the above example, the planning hazard ratio (HR) for survival rates of 60% vs. 70% at three years is 0.698. Based on Shuster,¹⁵ if each failure is considered to approximate an independent binomial event, which under the null hypothesis has a 50% probability of falling into each treatment and under the alternative a $\text{HR}/(1 + \text{HR}) = 0.698/1.698 = 0.411$ probability of being in the experimental arm vs. a 0.589 probability of being in the control arm, then from the results of Section 13.3, the study should accrue until 189 failures have occurred (nearly an identical number to the 196 expected failures derived from the piecewise exponential model).
5. The variance (the square of the right-hand side of Equation (13.3)) decreases more rapidly than $(1/X)$, the reciprocal of the accrual duration, because while the number of patients increases linearly, the longer time at risk increases the probability of death for each of the entrants. However,

for small increments in X , the rate of change is approximately proportional to the derivative of $(1/X)$, namely $-(1/X^2)$. This will be used to apply an inflation factor for sequential monitoring (Section 13.7).

6. The effect of ρ in the above example, where accrual is 2.33 years and minimum follow-up is 3 years, is relatively modest. Under exponentiality, 448 patients would be required ($\rho = 1$), the actual case $\rho = 0.5$, 490 patients would be required, while if $\rho = 0.0$, 561 patients would be needed. The effect of ρ is much more striking if accrual is slow. Under the same parameterization but with an accrual of 60 patients per year instead of 210, the patient requirements are respectively 353, 407, and 561 for $\rho = 1, 0.5$, and 0.0 , respectively.
7. Software for these calculations is available in Shuster,⁵ on a Windows platform. A Macintosh version exists but must invoke 16 bit architecture. [Appendix II](#) contains a SAS macro that also can be utilized for the calculations. The equal patient allocation of 50% to each arm is nearly optimal in terms of minimizing the smallest sample size, but the macro allows for both equal or unequal allocation.

13.7 SEQUENTIAL MONITORING BY THE O'BRIEN-FLEMING METHOD

In this section, from elementary asymptotic considerations of the Poisson processes obtained in exponential survival, one can use asymptotic Brownian motion to connect full sequential monitoring power to no monitoring power.

First, define a time parameter θ , $0 < \theta < 1$, with $\theta = 1$ being the maximum allowable time of completion, such that $\hat{\Delta}_\theta$ is asymptotically $N(\Delta, S^2/\theta)$. θ represents the ratio of variances of the estimate of effect size (final to interim analysis), and S^2 is the variance of the estimate of effect size at the planned final analysis ($\theta = 1$), calculated per Equation (13.3). From asymptotic considerations of the Poisson process (with the notation the same as the section on the exponential, except that the index θ was added to delineate the time scale) $\theta \hat{\Delta}_\theta$ is asymptotically a Brownian motion process with drift Δ and diffusion constant S^2 .

From first passage time considerations, (see Cox and Miller¹⁶ for example), on the basis of the rejection region for testing $\Delta = 0$ vs. $\Delta > 0$, the power function for the O'Brien-Fleming test that rejects the null hypothesis if at any time $\theta \hat{\Delta}_\theta > SZ_{\alpha/2}$, is

$$\begin{aligned} \Pi(\Delta, S, \alpha) &= \Phi[(\Delta/S) - Z_{\alpha/2}] \\ &+ \exp[(2\Delta Z_{\alpha/2})/S] \Phi[-(\Delta/S) - Z_{\alpha/2}] \end{aligned}$$

where Φ = standard normal CDF and Z_p is the upper $100p$ percentile of Φ . Note that $\Pi(0, S, \alpha) = \alpha$.

For no monitoring, the power function is defined by

$$\Pi_{\text{no}}(\Delta, S, \alpha) = \Phi[(\Delta/S) - Z_\alpha]$$

TABLE 13.1
Sensitivity of No Monitoring (None) vs. Continuous Monitoring by O’Brien–Fleming (OF). Δ/S represents the detectable effect size

		None	O-F
α	Power	Δ/S	Δ/S
0.010	0.80	3.168	3.238
0.025	0.80	2.801	2.881
0.050	0.80	2.486	2.576
0.010	0.90	3.608	3.688
0.025	0.90	3.242	3.3320
0.050	0.90	2.926	3.026

For example, if investigator #1 ran a study sensitive to an effect size $\Delta/S = 2.486$ and required no monitoring while investigator #2 ran a slightly larger study sensitive to an effect size of $\Delta/S = 2.576$ but used continuous O’Brien–Fleming bounds for sequential monitoring, the two studies would both have 80% power at $\alpha = 0.05$. There is almost no penalty for this type of monitoring, even if indeed the study runs to completion.

Because as remarked in Section 13.6.2, item 5, increasing the accrual duration slightly means that the variance changes proportional to the change in the reciprocal of the accrual duration, an approximate measure for the increase in accrual mandated by continuous O’Brien monitoring is approximately the square of the ratio of the entries in Table 13.1. The inflation factors would be 4% ($\alpha = 0.01$), 6% ($\alpha = 0.025$) and 7% ($\alpha = 0.05$).

Group sequential monitoring by the O’Brien–Fleming method would require a smaller inflation factor for the maximum sample size because it will fall between no monitoring and continuous monitoring. The software package EaST¹⁷ does not handle nonexponential data but can be used to derive an approximate inflation factor for true group sequential designs. But experience has shown it to be not much smaller than those derived for continuous monitoring.

In the example above, where 490 patients were needed for a trial conducted without sequential monitoring, an additional 7% would bring the necessary accrual to 524 (34 more entrants) if the trial were to be sequentially monitored by the O’Brien–Fleming method. This represents an accrual duration of at most 2.49 years with monitoring vs. 2.33 years without monitoring.

13.8 EVALUATION OF THE INFORMATION FRACTION

It is of interest to note that for the piecewise exponential model one can derive the information fraction as the ratio of the final to interim variance as in the strict definition (see Equation (13.3)). All one needs to compute is the expected number of failures at an interim analysis. Others use the ratio of expected failures (interim to final), while still others use the ratio of actual failures to total expected. Although the use of variance ratios Equation (13.3) appears to be quite different from the ratio of expected failures when the hazard ratios are not close to one, the fact is that for

studies planned with survival differences of the order of 15% or less where the expected total failures is 50 or higher, the two are almost identical.

In the numerical example above where planning accrual is 210 per year, planning difference is a 10% improvement in 3 year survival from 60% (control) to 70% (experimental) and monitoring is handled by a continuous O'Brien–Fleming approach, it was concluded that 2.49 years of accrual plus 3 years of minimal follow-up (total maximum duration of 5.49 years) were needed. The information fractions θ for calendar times 1.0 years, 1.5 years, 2.0 years, 2.5 years, 3.0 years, 3.5 years, 4.0 years, and 5.49 years are, respectively, for Equation (13.3) and for expected failures ratios shown in []: 1.0 years 0.0685 [0.0684]; 1.5 years 0.1504 [0.1502]; 2.0 years 0.2611 [0.2607]; 2.5 years 0.3985 [0.3979]; 3.0 years 0.5424 [0.5418]; 3.5 years 0.6704 [0.6698]; 4.0 years 0.7785 [0.7780]; 5.49 years 100% [100%]. These numbers are impressively close, despite the fact that under the alternate hypothesis the planning hazard ratio (experimental to control) is 0.698, hardly a local alternative to the null value of 1.00.

As noted above, the most robust concept is to use actual failures and terminate at a maximum of the expected number, which can be shown to be 211 for this continuous O'Brien–Fleming design.

13.9 MULTIPLE TREATMENTS AND STRATIFICATION

These methods can be applied to pairwise comparisons in multiple treatment trials. If one wishes to correct for multiple comparisons (a controversial issue), then one should use a corrected level of α but keep the original power. Note that the accrual rate would be computed for the pair of treatments being compared. For example, if the study is a three-armed comparison and accrual is estimated to be 300 patients per year, 200 per year would be accrued to each pairwise comparison. The definition of power applies to each pairwise comparison. In most applications, it is the author's opinion that the planning α level should not be altered for multiple comparisons. This is because the inference about A vs. B should be the same for the same data, whether or not there was a third arm C. In other words, had a hypothetical trial of only A vs. B been run and accrued the same data, one could reach a different conclusion if the actual analysis had been corrected for the number of pairwise comparisons in the original trial.

In general, the use of stratification on a limited scale can increase the precision of the comparison. However, the improvement over a completely randomized design is difficult to quantify. Hence, the nonstratified plan represents a conservative estimate of patient needs when the design and analysis are in fact stratified.

If a 2×2 factorial study is conducted, this could be analyzed as a stratified logrank test, stratifying for the concomitant treatment. Interventions should be carefully selected to minimize the potential for qualitative interaction, a situation where the superior treatment depends upon which of the concomitant treatments is employed. If the assumption of no major interaction in the hazard ratios is reasonable, a study planned as if it were a two-treatment, nonstratified study will generally yield a sample size estimate slightly larger (but unquantifiably so) than needed in the stratified analysis presuming proportional hazards within the strata hold. If a

qualitative interaction is indeed anticipated, then the study should be designed as a four-treatment trial for the purposes of patient requirements.

13.10 COMPETING LOSSES

In general, competing losses, patients censored for reasons other than being alive at the time of the analysis, are problematic unless they are treatment-uninformative (the reason for the loss is presumed to be independent of the treatment assignment and, at least conceptually, unrelated to the patient's prognosis). While it is possible to build in these competing losses in a quantified way for sample size purposes (See Rubinstein et al.),¹⁰ a conservative approach preferred by this author is to utilize a second inflation factor for competing losses. For example, if $L = 0.10$ (10%) are expected to be lost to competing reasons, the sample size would be inflated by dividing the initial sample size calculation by $(1 - L) = 0.9$ to obtain a final sample size.

13.11 PRACTICAL CONCLUSIONS

The methods proposed herein allow a planner to think in terms of fixed term survival rather than hazard rates or hazard ratios. The use of a fixed number of failures in implementation also allows for simplistic estimates of the information fraction as the failures to date divided by the total failures that would occur if the study is run to its completion.

Irrespective of the proportional hazards assumption, the actual test (logrank or Cox) is valid for testing the null hypothesis that the survival curves are identical. However, the sample size calculation and power are sensitive to violations of the proportional hazards assumption and especially to crossing hazards. If a superior treatment is associated with a high propensity for early deaths, an early significance favoring the wrong treatment may emerge causing the study to be stopped early, reaching the incorrect conclusion. For example, if at an early interim analysis, one treatment (e.g., bone marrow transplant) is more toxic and appears to have inferior outcome when compared to the other (chemotherapy alone), it might be very tempting but unwise to close the study for lack of efficacy. While sequential analysis is often an ethical necessity, it is also an ethical dilemma because statisticians are really being asked to forecast the future where they have little reliable information to work with.

It is recommended that studies where proportional hazards are considered to be a reasonable assumption be planned as if the piecewise exponential assumptions hold. If the plan is to monitor the study by the O'Brien–Fleming method, whether it is group sequential or pure sequential, apply the small inflation factor by taking the square of the ratio of the entry in the Δ/S columns per Table 13.1. Next, apply Equation (13.3) using expected values in Equations (13.4), (13.5), and (13.6) for the piecewise model to obtain the expected number of failures to be seen in the trial. Conduct the trial as planned until this many failures occur or until the study is halted for significance by crossing an O'Brien–Fleming bound where the information fraction is calculated as actual failures/maximum number of failures that would occur at the planned final analysis.

One possible recommendation for survival trials where there is concern about the model assumptions is to conduct accrual in stages. For example, it might be prudent to accrue patients onto a trial A for a period of 1 year and, irrespective of any interim results, close the trial temporarily and begin accrual on a new trial B for a period of 1 year. A decision as to whether to continue accrual to trial A for another year or begin yet another trial C for a year could be made with more mature data than otherwise possible. This leapfrog approach would continue in 1 year increments. This process might slow completion of trial A but also might speed completion of trials B and C. In addition, for safety purposes, there would be a much smaller disparity between calendar time and information time. In the numerical example initiated in Section 13.6 for the O'Brien–Fleming design, the information fraction at 1 year (about 40% of accrual) would be only 7%, while the information fraction at 2 years (about 80% of accrual completed) would be 26% (19% coming from the first year of accrual plus only 7% coming from the second year of accrual). In fact, if one had a 1 year “time out” for accrual but ran the second stage for 1.2 years analyzing the data 3.0 years after the end of accrual, the study would have the same power as the 2.49 year accrual study (both with continuous O'Brien–Fleming bounds). Completion would occur after 6.2 years instead of 5.49 years presuming the study ran to completion. Offsetting this difference, trial B would be completed earlier by the leapfrog approach by about 0.8 years. The slowdown generated by the leapfrog approach would also allow the analyst to have a better handle on model diagnostics while fewer patients were at risk.

ACKNOWLEDGMENT

This work was partially supported by Grant 29139 from the National Cancer Institute.

REFERENCES

1. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J. American Stat. Assoc.* 1958;53: 457–481.
2. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979;35: 549–556.
3. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *J. Royal Stat. Soc.* 1972;A135: 185–206.
4. Cox DR. Regression models and life tables (with discussion). *J. Royal Statist. Soc.* 1972;B34: 187–220.
5. Shuster JJ. *Practical Handbook of Sample Size Guidelines for Clinical Trials*. Boca Raton, FL: CRC Press, 1992.
6. Bernstein D, Lagakos SW. Sample size and power determination for stratified clinical trials. *J. Stat. Comput. Simul.* 1987;8: 65–73.
7. 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: 15–29.
8. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clin. Trials* 1981;2: 93–113.
9. Morgan TM. Planning the duration of accrual and follow-up for clinical trials. *J. Chronic Dis.* 1985;38: 1009–1018.

10. Rubinstein LJ, 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.
11. Schoenfeld DA. Sample size formula for the proportional-hazards regression model. *Biometrics* 1983;39: 499–503.
12. Haperin J, Brown BW. Designing clinical trials with arbitrary specification of survival functions and for the log rank test or generalized Wilcoxon test. *Controlled Clin. Trials* 1987;8: 177–189.
13. Cantor AB. Power estimation for rank tests using censored data: conditional and unconditional. *Controlled Clin. Trials* 1991;12: 462–473.
14. Henderson WG, Fisher SG, Weber L, Hammermeister KE, Sethi G. Conditional power for arbitrary survival curves to decide whether to extend a clinical trial. *Controlled Clin. Trials* 1991;12: 304–313.
15. Shuster JJ. Fixing the number of events in large comparative trials with low event rates: a binomial approach. *Controlled Clin. Trials* 1993;14: 198–208.
16. Cox DR, Miller HD. *The theory of Stochastic Processes*. London: Methuen, 1965.
17. Mehta C. EaST: (Early Stopping in Clinical Trials), Cambridge, MA: CyTEL Software Corporation.

APPENDIX I:

LOGRANK VS. EXPONENTIAL ESTIMATION–LOCAL OPTIMALITY

For treatment i and day j , let N_{ij} and F_{ij} represent the number at risk at the start of the day and number of failures on that day, respectively.

For the logrank test, the contribution to the observed minus expected for day j (to the numerator of the statistic, where the denominator is simply the standard error of the numerator) is:

$$F_{1j} - N_{1j}(F_{1j} + F_{2j})/(N_{1j} + N_{2j}) = \{(F_{1j}/N_{1j}) - (F_{2j}/N_{2j})\}/(N_{1j}^{-1} + N_{2j}^{-1})$$

The day j estimate of each hazard, F_{ij}/N_{ij} , is weighted inversely proportional to $(N_{1j}^{-1} + N_{2j}^{-1})$, that is proportional to $N_{1j}N_{2j}/(N_{1j} + N_{2j})$. The weight is zero if either group has no patient at risk starting day j .

For the exponential test, the contributions to the estimates of the hazards for day j are: $(N_{ij}/N_i)(F_{ij}/N_{ij})$, where N_i is the total days on test for treatment i . Under the null hypothesis and exponentiality, both weight the information inversely proportional to the variance, and as such, both tests are locally optimal.

Using the approximation that the N_{ij} are fixed rather than random, a relative efficiency measure can easily be obtained for nonlocal alternatives, and in practice, the relative efficiency for studies planned as above with 30 or more failures will generally be only trivially less than 1.00 (typically 0.99 or higher). Note that two conditions help the logrank test. First, if the ratio of N_{1j}/N_{2j} remains fairly constant over time $j = 1, 2$, etc., until relatively few are at risk, the weights will be very similar except when both become very small. Second, stratified estimation is robust against modest departures from the optimal allocation, which is represented by the exponential weights.

Comment: The mathematical treatment of this chapter is somewhat unconventional in that it utilizes the difference in hazards for the exponential rather than the hazard ratio or log hazard ratio. This enables a user to directly investigate the relative efficiency issue of the logrank to exponential test. On the negative side, by not using a transformation, the connection to expected failures is not as direct as it would be under the transform to a hazard ratio.

Comment: SAS (Statistical Analysis Systems) macros for the logrank and stratified logrank tests are included in [Appendix III](#).

APPENDIX II:

SAS MACRO FOR SAMPLE SIZE AND ACCRUAL DURATION

Note that this macro can handle unequal allocation of patients, something useful for planning studies for the prognostic significance of yes/no covariates.

Usage

```
%ssize(ddsetx,alloc,psi,y,r1,r2,side,alpha,pi,rho,lfu);
%ssize(a,alloc,psi,y,r1,r2,side,alpha,pi,rho,lfu);
```

ddsetx = user supplied name for data set containing the planning parameters. See third paragraph of Section VI.

Alloc = Fraction allocated to control group (Use .5 for 1-1 randomized trials)

Psi = Annual Accrual Rate

Y = Minimum Follow-up

R1 = Control Group planned Y-Year Survival

R2 = Experimental Group planned Y-Year Survival

side = 1 (one-sided) or 2(two-sided)

alpha = size of type I error

pi = Power

rho = Average Post:pre Y-Year hazard ratio

lfu = Fraction expected to be lost to follow-up.

```
%MACRO ssize(ddsetx,alloc,psi,y,r1,r2,side,alpha,pi,rho,lfu);
options ps = 60 ls = 75;
OPTIONS NOSOURCE NONOTES;
DATA DDSETX;set &ddsetx;
alloc = &alloc;psi = &psi;y = &y;r1 = &r1;r2 = &r2;alpha = &alpha;
pi = &pi;rho = &rho;lfu = &lfu;
lam1 = -log(r1)/y;
lam2 = -log(r2)/y;
del = abs(lam1-lam2);
za = -probit(alpha/side);
zb = probit(pi);
s = del/(za+zb);
```

```

x = 0;inc = 1;
aa:x = x + inc;
q1 = r1*(1-exp(-rho*lam1*x))/(rho*lam1*x);
q2 = r2*(1-exp(-rho*lam2*x))/(rho*lam2*x);
p1 = 1-q1;
p2 = 1-q2;
v1 = ((lam1**2)/(alloc*psi*x*p1));
v2 = ((lam2**2)/((1-alloc)*psi*x*p2));
s2 = sqrt(v1 + v2);
if s2>s and inc>.9 then goto aa;
if inc>.9 then do;inc = .01;x = x-1;goto aa;end;
if s2>s then goto aa;
n = int(.999 + psi*x);
na = n/(1-lfu);na = int(na + .999);
ex_fail = psi*x*(alloc*p1 + (1-alloc)*p2);
data ddsetx;set ddsetx;
label alloc = 'Allocated to Control'
rho = 'Post to Pre y-Year Hazard'
psi = 'Accrual Rate'
y = 'Minimum Follow-up (MF)'
r1 = 'Planned control Survival at MF'
r2 = 'Planned Experimental Survival at MF'
side = 'Sidedness'
alpha = 'P-Value'
pi = 'power'
lfu = 'Loss to Follow-up Rate'
x = 'Accrual Duration'
n = 'Sample size, no losses'
na = 'Sample size with losses'
ex_fail = 'Expected Failures';
proc print label;
var alloc psi y r1 r2 side alpha pi rho lfu x n na ex_fail;
%mend;

```

To produce the example in Section 13.6, this program was run.

```

data ddsetx;
input alloc psi y r1 r2 side alpha pi rho lfu;
cards;
.5 210 3 .6 .7 1 .05 .8 1 .1
.5 210 3 .6 .7 1 .05 .8 .5 .1
.5 210 3 .6 .7 1 .05 .8 .000001 .1
.5 60 3 .6 .7 1 .05 .8 1 .1
.5 60 3 .6 .7 1 .05 .8 .5 .1
.5 60 3 .6 .7 1 .05 .8 .000001 .1
%ssize(ddsetx,alloc, psi,y,r1,r2,side,alpha,pi,rho,lfu);

```

APPENDIX III:

SAS MACRO FOR LOGRANK AND STRATIFIED LOGRANK TESTS

Sample usage:

```
%logrank(ddset,time,event,class)
```

```
%logrank(a,x1,x2,treat_no)
```


for the unstratified logrank test
`%logstr(ddset,time,event,class,stratum)`
`%logstr(a,x1,x2,treat_no, gender)`
 for the stratified logrank test

time, event, class, and stratum are user supplied names in the user supplied dataset called ddset.

ddset = name of dataset needing analysis.

time = time variable in data set ddset (time<0 not allowed)

event = survival variable in data set ddset event = 0 (alive) or event = 1 (died)

class = categorical variable in ddset identifying treatment group.

Stratum = categorical variable in ddset defining stratum for stratified logrank test.

```
%MACRO LOGRANK (DDSET, TIME, EVENT, CLASS) ;
OPTIONS NOSOURCE NONOTES
MISSING = ' ' ;
DATA DDSETX; SET &DDSET; STRATUM = 1;
KEEP &TIME &EVENT &CLASS STRATUM;
IF &TIME<0 OR &EVENT = . OR &CLASS = ' ' THEN DELETE;
TITLE3 'LOGRANK TEST BY SUM (O-E) **2/E';
TITLE4 'OBSERVED AND EXPECTED BY GROUP';
TITLE5 "TIME = &TIME EVENT = &EVENT COMPARED = &CLASS";
DATA DDSETX; SET DDSETX;
EVT = -&EVENT;
PROC SORT; BY STRATUM &TIME EVT;
DATA DDSETX; SET DDSETX; PROC MEANS NOPRINT; BY STRATUM;
VAR &EVENT; OUTPUT OUT = ANUMB N = N MAX = EVMAX;
DATA DDSETX; MERGE DDSETX ANUMB; BY STRATUM;
DATA DDSETX; SET DDSETX; EVENT = &EVENT;
IF EVMAX = 2 THEN EVENT = EVENT-1;
IF OLDSTRAT = STRATUM THEN GOTO AA;
NN = N; NOWRISK = N;
FAIL = EVENT; GOTO BB;
AA: IF EVENT = 0 OR (&TIME NE OLDTIME) THEN NOWRISK = NN;
EX = FAIL/NOWRISK;
IF (OLDTIME = &TIME)
AND EVENT = 1 THEN FAIL = FAIL+1; ELSE FAIL = EVENT;
BB: EX = FAIL/NOWRISK;
NN = NN-1; OUTPUT;
OLDSTRAT = STRATUM;
OLDTIME = &TIME;
RETAIN NN OLDSTRAT OLDTIME NOWRISK FAIL;
DATA DDSETX; SET DDSETX;
PROC MEANS NOPRINT; BY STRATUM &TIME EVT;
VAR EX; OUTPUT OUT = ANUMB MAX = EXD;
DATA DDSETX; MERGE DDSETX ANUMB; BY STRATUM &TIME EVT;
DATA DDSETX; SET DDSETX;
```

```

IF OLDS = STRATUM THEN GOTO AA;EXPECTED = EXD;GOTO BB;
AA:IF (OLDT NE &TIME OR OLDE = 0) AND EVENT = 1 THEN EXPECTED =
EXPECTED+EXD;
BB:OLDS = STRATUM;OLDT = &TIME;OLDE = EVENT;
OUTPUT;
RETAIN OLDS OLDT OLDE EXPECTED;
DATA DDSETX;SET DDSETX;
PROC SORT;BY STRATUM &CLASS;
PROC MEANS SUM N NOPRINT;BY STRATUM &CLASS;VAR EVENT EXPECTED;
OUTPUT OUT = ANUMB SUM = OBS EX N = N N1;
DATA ANUMB;SET ANUMB;Y = (OBS-EX)*(OBS-EX);OBSERVED = OBS;
EXPECTED = EX;CHISQ = 0;IF EX>0 THEN CHISQ = Y/EX;IF EX>0 THEN OE =
OBS/EX;
RATIO_OE = OE;

PROC MEANS NOPRINT;BY STRATUM;VAR CHISQ;
OUTPUT OUT = BNUMB SUM = CHISQ N = DF;
DATA BNUMB;SET BNUMB;DF = DF-1;IF DF<.2 THEN GOTO CC;
P_VALUE = 1-PROBCHI(CHISQ,DF);CC;;
DATA ANUMB;MERGE ANUMB BNUMB;BY STRATUM;
OVERALL = 'NO ';IF STRATUM = ' ' THEN OVERALL = 'YES';
PROC SORT;BY STRATUM OVERALL;
DATA ANUMB;SET ANUMB;BY STRATUM;
IF FIRST.STRATUM THEN GOTO AA;
CHISQ = .;DF = .;P_VALUE = .;AA;;
PROC PRINT;
ID &CLASS;VAR N OBSERVED EXPECTED RATIO_OE
CHISQ DF P_VALUE;
PROC SORT;BY STRATUM;TITLE ' ';
MISSING A B C D E F G H I J K L M N O P Q R S T U V W X Y Z;
OPTIONS SOURCE NOTES;
DATA &DDSET;SET &DDSET;
%MEND;

%MACRO LOGSTR(DDSET,TIME,EVENT,CLASS,STRATUM);
OPTIONS NOSOURCE NONOTES
MISSING = ' ';
DATA DDSETX;SET &DDSET;
KEEP &TIME &EVENT &CLASS &STRATUM;
IF &TIME<0 OR &EVENT = . OR &CLASS = ' ' THEN DELETE;
TITLE3 'STRATIFIED LOGRANK TESTS BY SUM (O-E)**2/E';
TITLE4 'OBSERVED AND EXPECTED BY STRATUM AND GROUP';
TITLE5 "TIME = &TIME EVENT = &EVENT STRATUM = &STRATUM COMPARED = &CLASS";
DATA DDSETX;SET DDSETX;
EVT = -&EVENT;
PROC SORT;BY &STRATUM &TIME EVT;
PROC MEANS NOPRINT;
VAR &EVENT;OUTPUT OUT = BNUMB MAX = EVMAX;
DATA DDSETX;SET DDSETX;PROC MEANS NOPRINT;BY &STRATUM;
VAR &EVENT;OUTPUT OUT = ANUMB N = N ;
DATA ANUMB;SET ANUMB;TT = 1;PROC SORT;BY TT;
DATA BNUMB;SET BNUMB;TT = 1;PROC SORT;BY TT;
DATA ANUMB;MERGE ANUMB BNUMB;BY TT;
PROC SORT;BY &STRATUM ;

```

```

DATA DDSETX;MERGE DDSETX ANUMB;BY &STRATUM;
DATA DDSETX;SET DDSETX;DROP TT;EVENT = &EVENT;
IF EVMAX = 2 THEN EVENT = EVENT-1;
IF OLDSTRAT = &STRATUM THEN GOTO AA;
NN = N;NOWRISK = N;
FAIL = EVENT;GOTO BB;
AA:IF EVENT = 0 OR (&TIME NE OLDTIME) THEN NOWRISK = NN;
EX = FAIL/NOWRISK;
IF (OLDTIME = &TIME)
AND EVENT = 1 THEN FAIL = FAIL+1;ELSE FAIL = EVENT;
BB:EX = FAIL/NOWRISK;
NN = NN-1;OUTPUT;
OLDSTRAT = &STRATUM;
OLDTIME = &TIME;
RETAIN NN OLDSTRAT OLDTIME NOWRISK FAIL;
DATA DDSETX;SET DDSETX;
PROC MEANS NOPRINT;BY &STRATUM &TIME EVT;
VAR EX;OUTPUT OUT = ANUMB MAX = EXD;
DATA DDSETX;MERGE DDSETX ANUMB;BY &STRATUM &TIME EVT;
DATA DDSETX;SET DDSETX;
IF OLDS = &STRATUM THEN GOTO AA;EXPECTED = EXD;GOTO BB;
AA:IF (OLDT NE &TIME OR OLDE = 0) AND EVENT = 1 THEN EXPECTED =
EXPECTED+EXD;
BB:OLDS = &STRATUM;OLDT = &TIME;OLDE = EVENT;
OUTPUT;
RETAIN OLDS OLDT OLDE EXPECTED;
DATA DDSETX;SET DDSETX;
PROC SORT;BY &STRATUM &CLASS;
PROC MEANS SUM N NOPRINT;BY &STRATUM &CLASS;VAR EVENT EXPECTED;
OUTPUT OUT = ANUMB SUM = OBS EX N = N N1;
DATA ANUMB;SET ANUMB;
OVERALL = 'INCLUDE';IF &STRATUM = ' ' THEN OVERALL = 'EXCLUDE';
DATA CNUMB;SET ANUMB;IF &STRATUM = ' ' THEN DELETE;
PROC SORT;BY &CLASS;
PROC MEANS SUM N NOPRINT;BY &CLASS;VAR OBS EX N;
OUTPUT OUT = BNUMB SUM = OBS EX N;
DATA BNUMB;SET BNUMB;OVERALL = 'TOTAL ' ;
DATA ANUMB;SET ANUMB BNUMB;PROC SORT;BY OVERALL &STRATUM;
DATA ANUMB;SET ANUMB;Y = (OBS-EX)*(OBS-EX);OBSERVED = OBS;
EXPECTED = EX;CHISQ = 0;IF EX>0 THEN CHISQ = Y/EX;IF EX>0 THEN OE =
OBS/EX;
RATIO_OE = OE;

PROC MEANS NOPRINT;BY OVERALL &STRATUM;VAR CHISQ;
OUTPUT OUT = BNUMB SUM = CHISQ N = DF;
DATA BNUMB;SET BNUMB;DF = DF-1;IF DF<.2 THEN GOTO CC;
P_VALUE = 1-PROBCHI(CHISQ,DF);CC:;
DATA ANUMB;MERGE ANUMB BNUMB;BY OVERALL &STRATUM;
DATA ANUMB;SET ANUMB;
IF Y = . THEN GOTO AA;
IF OLDS NE &STRATUM OR OLDO NE OVERALL THEN GOTO AA;
CHISQ = .;DF = .;P_VALUE = .;
AA:;Y = 1;OUTPUT;OLDS = &STRATUM;OLDO = OVERALL;RETAIN OLDS OLDO Y;
PROC PRINT;BY OVERALL &STRATUM;
ID &CLASS;VAR N OBSERVED EXPECTED RATIO_OE

```

```
CHISQ DF P_VALUE;  
PROC SORT;BY &STRATUM;TITLE ' ';  
MISSING A B C D E F G H I J K L M N O P Q R S T U V W X Y Z;  
OPTIONS SOURCE NOTES;  
DATA &DDSET;SET &DDSET;  
%MEND;
```

14 Early Stopping of Cancer Clinical Trials

*James J. Dignam, John Bryant, and
H. Samuel Wiewand*

CONTENTS

14.1	Introduction	227
14.2	Group Sequential Monitoring Rules	228
14.2.1	General Principles	228
14.2.2	Symmetric Two-sided Group Sequential Stopping Boundaries	229
14.2.3	Asymmetric Stopping Boundaries	230
14.3	Conditional Power Methods.....	232
14.3.1	Stochastic Curtailment	232
14.3.2	Predictive Power and Current Data Methods.....	233
14.4	A Bayesian Approach to Assessing Early Termination for No Benefit.....	234
14.5	Examples	235
14.5.1	A Trial Stopped Early for No Benefit	235
14.5.2	The Effect of Two Easily Applied Rules in the Adjuvant and Advanced Disease Setting	238
14.5.2.1	Scenario 1	240
14.5.2.2	Scenario 2	241
14.6	Summary	241
	References	244

14.1 INTRODUCTION

Most cancer clinical trials use formal statistical monitoring rules to serve as guidelines for possible early termination. Such rules provide for the possibility of early stopping in response to positive trends that are sufficiently strong to establish the treatment differences the clinical trial was designed to detect. At the same time, they guard against prematurely terminating a trial on the basis of initial positive results that may not be maintained with additional follow-up.

We may also consider stopping a trial before its scheduled end point if current trends in the data indicate that eventual positive findings are unlikely. For example, consider a trial comparing a new treatment to an established control regimen. Early

termination for negative results may be called for if the data to date are sufficient to rule out the possibility of improvements in efficacy that are large enough to be clinically relevant. Alternatively, it may have become clear that study accrual, drug compliance, follow-up compliance, or other factors have rendered the study incapable of discovering a difference, whether or not one exists.

In this chapter we discuss methods for early stopping of cancer clinical trials. We will focus in particular on stopping for *futility*, i.e., situations where evidence suggests that differences in efficacy between treatments will not be demonstrated, as this aspect of trial monitoring has historically received less attention. For concreteness, we will restrict our attention to randomized clinical trials designed to compare two treatments using survival (or slightly more generally, time to some event) as the primary criterion. However, the methods we will discuss may be extended to other trial designs and efficacy criteria. In most applications, one treatment represents an established regimen for the disease and patient population in question, whereas the second is an experimental regimen to be tested by randomized comparison with this standard.

In this chapter, we first describe group sequential approaches to trial monitoring and outline a general framework for designing group sequential monitoring rules. We then discuss the application of asymmetric monitoring boundaries to clinical trials in situations where it is appropriate to plan for the possibility of early termination in the face of negative results. Next, we consider various approaches to assessing futility in ongoing trials including predictive methods such as stochastic curtailment. We then briefly examine Bayesian methods for trial monitoring and early stopping. National Surgical Adjuvant Breast and Bowel Project (NSABP) Protocol B-14 is presented as a detailed example illustrating the use of stochastic curtailment calculations and Bayesian methods. A second example is also given to illustrate the use of a common asymmetric monitoring plan adapted for use in Southwest Oncology Group (SWOG) Protocol SWOG-8738. This approach is compared to a slight modification of a monitoring rule proposed earlier by Wieand et al. We conclude with a discussion of considerations relevant to the choice of a monitoring plan.

14.2 GROUP SEQUENTIAL MONITORING RULES

14.2.1 GENERAL PRINCIPLES

The most common statistical monitoring rules are based on *group sequential* procedures. Consider a clinical trial designed to compare two treatments, A and B, using survival (or some other well-defined event) as the primary end point. The relative effectiveness of the two treatments can be summarized by the parameter $\delta = \ln(\lambda_B(t)/\lambda_A(t))$, the logarithm of the ratio of hazard rates $\lambda_B(t)$ and $\lambda_A(t)$. We assume that this ratio is independent of time t . Thus, the hypothesis that the two treatments are equivalent is $H_0: \delta = 0$, while values of $\delta > 0$ indicate the superiority of A to B and values of $\delta < 0$ indicate the superiority of B to A.

Suppose patients are accrued and assigned at random to receive either treatment A or B. In a group sequential test of H_0 , information is allowed to accumulate over

time; at specified intervals an interim analysis is performed, and a decision is made whether to continue with the accumulation of additional information or to stop and make some decision based on the information collected to date. Typically, the accumulation of information is quantified by the total number of events, and the comparison of treatments is based on the logrank statistic.

A large number of group sequential procedures have been proposed in this setting. Most fall into a common general framework that we now describe: For $k = 1, 2, \dots, K-1$, an interim analysis is scheduled to take place after m_k total events have occurred (summed over both treatment arms), and a final analysis is scheduled to occur after the m_K th event. Let L_k denote the logrank statistic computed at the k th analysis; let V_k denote its variance; and let Z_k represent the corresponding standardized statistic $Z_k = L_k/\sqrt{V_k}$. For each $k = 1, 2, \dots, K$, the real line R^1 is partitioned into a continuation region C_k and a stopping region $S_k = R^1 - C_k$; if $Z_k \in C_k$, we continue to the $(k + 1)$ st analysis, but if $Z_k \in S_k$, we stop after the k th analysis. The stopping region for the K th analysis is the entire real line, $S_K = R^1$.

Define $t_k = m_k/m_K$, so that t_k represents the fraction of the total information available at the k th analysis, and let $W_k = Z_k \cdot \sqrt{t_k}$, $k = 1, 2, \dots, K$. Under appropriate conditions (roughly, sequential entry, randomized treatment assignment, loss-to-follow-up independent of entry time and treatment assignment), the W_k behave asymptotically like Brownian motion. Defining $\Delta t_k = t_k - t_{k-1}$ and $\eta = \delta \cdot \sqrt{m_K}/2$, the increments $W_k - W_{k-1}$ are approximately uncorrelated normal random variables with means $\eta \cdot \Delta t_k$ and variances Δt_k .¹⁻³ This result permits the extension of sequential methods based on evolving sums of independent normal variates to the survival setting. In particular, the recursive integration scheme of Armitage et al.⁴ may be used to compute the density

$$f_k(w; \eta) = d\Pr\{\tau \geq k, W_k \leq w; \eta\}/dw$$

where τ represents the number of the terminal analysis. For the first analysis time, $f_1(w; \eta) = \phi\{(w - \eta \cdot t_1)/\sqrt{t_1}\}/\sqrt{t_1}$, where $\phi\{\cdot\}$ represent the standard normal density, and for $k = 2, 3, \dots, K$,

$$f_k(w; \eta) = \int_{C_{k-1}} f_{k-1}(y; \eta) \cdot [\{(w - y - \eta \cdot \Delta t_k)/\sqrt{\Delta t_k}\}/\sqrt{\Delta t_k}] dy \quad (14.1)$$

From this result, all operating characteristics of the group sequential procedure, such as size, power, and stopping probabilities, may be obtained.

14.2.2 SYMMETRIC TWO-SIDED GROUP SEQUENTIAL STOPPING BOUNDARIES

In cases where a two-sided symmetric test of the hypothesis $H_0: \delta = 0$ is appropriate, the continuation regions are of the form $C_k = \{Z_k: -b_k \leq Z_k \leq b_k\}$, $k = 1, 2, \dots, K-1$. If $Z_k < -b_k$ at the k th analysis, we reject H_0 in favor of $H_A: \delta < 0$, whereas if $Z_k > b_k$, H_0 is rejected in favor of $H_A: \delta > 0$. If testing continues to the K th and final analysis, a similar decision rule is applied except that if $-b_K \leq Z_K \leq b_K$, we accept

H_0 rather than continuing to an additional analysis. The b_k are chosen to maintain a desired experiment-wise type I error rate $\Pr\{\text{Reject } H_0 \mid H_0\} = \alpha$, and the maximum duration m_K (in terms of number of events) of the trial is selected to achieve power $1 - \beta$ against a specified alternative $H_A: \delta = \delta_A$ by determining that value of η which yields $\Pr\{\text{Reject } H_0 \mid \eta\} = 1 - \beta$ and then setting

$$m_K = 4 \cdot \eta^2 / \delta_A^2. \quad (14.2)$$

Early stopping rules proposed by Haybittle,⁵ Pocock,^{6,7} O'Brien and Fleming,⁸ Wang and Tsatis,⁹ and Fleming, Harrington, and O'Brien¹⁰ all fit into this general framework. In the method by Haybittle, a constant large critical value is used for analyses $k = 1, 2, \dots, K-1$, and the final analysis is performed using a critical value corresponding to the desired overall type I error level. For a moderate number of analyses (say, $K = 5$) and a large critical value, such as $Z = 3.0$, if one wishes to obtain an overall 0.05 level procedure, the method can be shown to achieve nearly the desired type I error rate despite no adjustment to the final test boundary value.⁵ To obtain the final critical value that would yield precisely the desired α level overall, Equation (14.1) can be used. The Pocock bounds are obtained by constraining the z -critical values to be identical for each k : $b_k = \text{constant}$, $k = 1, 2, \dots, K$. For the O'Brien-Fleming procedure, the W -critical values are constant, so that $b_k = \text{constant} / \sqrt{t_k}$. Wang and Tsatis⁹ boundaries have the form $b_k = \text{constant} \cdot t_k^{\Delta-1/2}$, where Δ is a specified constant. The boundaries of Fleming et al.¹⁰ retain the desirable property of the O'Brien-Fleming procedure that the nominal level of the K th analysis is nearly equal to α but avoid the extremely conservative nature of that procedure for small k when $K > 3$.

14.2.3 ASYMMETRIC STOPPING BOUNDARIES

Since most phase III trials compare a new therapy A to an accepted standard B, it may oftentimes be appropriate to consider one-sided hypothesis tests of $H_0: \delta = 0$ versus $H_A: \delta > 0$ and to make use of asymmetric continuation regions of the form $C_k = \{Z_k: a_k \leq Z_k \leq b_k\}$, or equivalently $C_k = \{W_k: A_k \leq W_k \leq B_k\}$ where $A_k = a_k \cdot \sqrt{t_k}$, $B_k = b_k \cdot \sqrt{t_k}$. Crossing the upper boundary results in rejection of H_0 in favor of H_A , whereas crossing the lower boundary results in trial termination in recognition that the new therapy is unlikely to be materially better than the accepted standard or that H_0 is unlikely to be rejected with further follow-up. The design of such asymmetric monitoring plans presents no significant new computational difficulties. After restricting the choice of a_k and b_k , $k = 1, 2, \dots, K$, to some desired class of boundaries, Equation (14.1) is used (generally in an iterative fashion) to fix both the size of the monitoring procedure and its power against a suitable alternative or set of alternatives. Equation (14.2) is used to determine the maximum duration of the trial in terms of observed events.

In this context, DeMets and Ware¹¹ proposed the use of asymmetric Pocock boundaries: the lower boundary points a_k are fixed at some specified value independent of k (the range $-2.0 \leq a_k \leq -0.5$ is tabled), and then a constant value for the b_k is determined by setting the type I error rate to α . A second suggestion was to

employ a test with boundaries that are motivated by their similarity to those of a sequential probability ratio test.¹² This procedure is most easily expressed in terms of its W -critical values, which are linear in information time:

$$A_k = -(Z'_L/\eta) + (\eta/2) \cdot t_k, B_k = (Z'_U/\eta) + (\eta/2) \cdot t_k, k = 1, 2, \dots, K \quad (14.3)$$

Here $Z'_L = \ln((1 - \alpha)/\beta)$, and Z'_U and η are chosen to satisfy type I and type II error requirements by iterative use of Equation (14.1). The maximum number of observed events is given by Equation (14.2), as before. In a subsequent publication¹³ DeMets and Ware recommend that the Wald-like lower boundary be retained but the upper boundary be replaced by an O'Brien–Fleming boundary $B_k \equiv B, k = 1, 2, \dots, K$. Although iteration is still required to determine η and B , the value of B is reasonably close to the symmetric O'Brien–Fleming bound at level 2α .

Whitehead and Stratton¹⁴ indicate how the sequential triangular test^{15,16} may be adapted to the group sequential setting in which K analyses will be carried out at equally spaced intervals of information time, $t_k = k/K, k = 1, 2, \dots, K$. Suppose first that it is desired to achieve type I error rate α and power $1 - \beta$ against the alternative $\delta = \delta_A$. The W -critical values are:

$$A_k = -Q + (3\eta/4) \cdot t_k, B_k = Q + (\eta/4) \cdot t_k, k = 1, 2, \dots, K$$

where η satisfies $\eta^2 + \{2.332/\sqrt{K}\} \cdot \eta - 8 \cdot \ln(1/2\alpha) = 0$, and $Q = 2 \cdot \ln(1/2\alpha)/\eta - 0.583/\sqrt{K}$. The maximum number of observed events required to achieve this is given by Equation (14.2). If instead one wishes to achieve a power of $1 - \beta \neq 1 - \alpha$ against the alternative $\delta = \delta_A$, the operating characteristic curve of the fixed sample size test satisfying $\Pr\{\text{Reject } H_0 \mid \delta = 0\} = \alpha, \Pr\{\text{Reject } H_0 \mid \delta = \delta_A\} = 1 - \beta$ may be used to determine an alternative δ'_A such that $\Pr\{\text{Reject } H_0 \mid \delta = \delta'_A\} = 1 - \alpha$. Then δ'_A should be used in place of δ_A in Equation (14.2). The adjustment factor of $0.583/\sqrt{K}$ in the formula for Q is an approximate correction for exact results that hold in the case of a purely sequential procedure. Slightly more accurate results may be obtained by iteratively determining η and Q to satisfy type I and type II error constraints via Equation (14.1). A related monitoring plan based solely on a lower futility boundary that does not permit early stopping for efficacy also has been proposed¹⁷. While such a rule would not likely be suitable for cancer treatment trials, it may be applicable in certain settings where a maximum sample size trial is desirable along with a provision for early futility stopping, such as cancer prevention trials based on surrogate endpoints.

The triangular test approximately minimizes the expected number of events at termination under the alternative $\delta = \delta'_A/2$. Jennison¹⁸ considers group sequential tests that minimize the expected number of events under various alternatives and presents parametric families of tests that are nearly optimal in this sense. These are specified in terms of spending functions similar to Lan and DeMets.¹⁹

The power boundaries of Wang and Tsatis⁹ may be adapted for use in testing hypotheses of the form $H_0: \delta = 0$ vs. $H_A: \delta = \delta_A$.^{20,21} W -critical values are of the form

$$A_k = -Q t_k^\Delta + \eta \cdot t_k, B_k = (\eta - Q) \cdot t_k^\Delta, k = 1, 2, \dots, K$$

where the constant Δ is specified by the trial designer; η and Q are determined iteratively to satisfy type I and type II error constraints using Equation (14.1). The maximum number of events required is given by Equation (14.2). $\Delta = 0$ corresponds essentially to a design using an upper O'Brien–Fleming bound to test $H_0: \delta = 0$, and a lower O'Brien–Fleming bound to test $H_A: \delta = \delta_A$. $\Delta = 1/2$ results in Pocock-like boundaries. In general, larger values of Δ correspond to a greater willingness to terminate at an earlier stage. Emerson and Fleming²⁰ compare the efficiencies of one-sided symmetric designs having power boundaries to the results of Jennison¹⁸ and conclude that the restriction to boundaries of this form results in negligible loss of efficiency. Pampallona and Tsiatis²¹ provide a comparison of the operating characteristics of asymmetric one-sided designs based on power boundaries with the designs proposed by DeMets and Ware.^{11,13} Both Emerson and Fleming²⁰ and Pampallona and Tsiatis²¹ also consider two-sided group sequential procedures that allow for the possibility of early stopping in favor of the null hypothesis. These procedures are similar in spirit to the double triangular test.¹⁴

Wieand, Schroeder, and O'Fallon²² proposed a method for early termination of trials when there appears to be no benefit after a substantive portion of total events have been observed, which is tantamount to adopting asymmetric boundaries after sufficient information has been obtained to guarantee high power against alternatives of interest. The method is an extension of earlier work by Ellenberg and Eisenberger²³ and Wieand and Therneau²⁴ and was first considered for multistage trials in advanced disease where patient outcomes are poor and there is likely to be substantial information regarding treatment efficacy while accrual is still underway. In its simplest form, the proposed rule calls for performing an interim analysis when one half of the required events have been observed. At that time, if the event rate on the experimental arm exceeds that on the control arm (e.g., hazard ratio ≤ 1.0), then termination of the trial should be considered. It can be shown that the adoption of this rule has essentially no effect on the size of a nominal 0.05-level test of equality of hazards and results in a loss of power of ≤ 0.02 for any alternative hypothesis indicating a treatment benefit, compared with a fixed sample size test of that same alternative at the scheduled definitive analysis.²² Similarly, if this rule is superimposed on symmetric two-sided boundaries by replacing the lower boundary a_k with 0 for any information time t_k greater than or equal to $1/2$ and the result is viewed as an asymmetric group sequential procedure testing a one-sided hypothesis, there is almost no change in the operating characteristics. In this implementation, the stopping rule calls for early termination if at any scheduled interim analysis at or after the halfway point the experimental treatment is observed to be no more efficacious than the control.

14.3 CONDITIONAL POWER METHODS

14.3.1 STOCHASTIC CURTAILMENT

A commonly applied predictive approach to early stopping makes use of the concept of stochastic curtailment.^{25–28} The stochastic curtailment approach requires a computation of *conditional power* functions, defined as

$$\gamma_i = \Pr(Z(1) \in R \mid D, H_i), \quad (14.4)$$

where $Z(1)$ represents a test statistic to be computed at the end of the trial, R is the rejection region of this test, D represents current data, and H_i denotes either the null hypothesis H_0 or an alternative hypothesis H_A . If this conditional probability is sufficiently large under H_0 , one may decide to stop and immediately reject H_0 . Conversely, if under a “realistic” alternative hypothesis H_A this probability is sufficiently small, or equivalently, if $1 - \gamma_A = \Pr(Z(1) \notin R | D, H_A)$ is sufficiently large, we may decide that continuation is futile because H_0 ultimately will not be rejected regardless of further observations. This is the case of interest when considering early stopping for futility. In an example presented later, we condition on various alternatives in favor of the treatment to assess the potential for a trial to reverse from early interim analyses results unexpectedly favoring the control group.

In Section 14.2 it was noted that the normalized logrank statistics asymptotically behave like Brownian motion. This provides an easy way to compute conditional power over a range of alternatives:^{27,28}

$$C(t) = 1 - \Phi \left(\frac{(Z_\alpha - Z(t)\sqrt{t} - \eta(1 - t))}{\sqrt{1 - t}} \right). \quad (14.5)$$

In Equation (14.5), $\Phi(\cdot)$ is the standard normal distribution function; t is the fraction of total events for definitive analysis that have occurred to date (so-called information time; this was defined for prespecified increments as $t_k = m_k/m_K$ in Section 14.2); $Z(t)$ is the current standard normal variate associated with the logrank test; Z_α is the critical value against which the final test statistic $Z(1)$ is to be compared; and η is defined in Section 14.2.

14.3.2 PREDICTIVE POWER AND CURRENT DATA METHODS

Stochastic curtailment has been criticized on the basis that it requires conditioning on the current data and at the same time an alternative hypothesis that may be unlikely to have given rise to that data, and in any case, because some H_A must be specified, the method always depends on unknown information at the time of the decision.²⁹ One might instead consider methods that take an unconditional predictive approach in assessing the consequences of continuing the trial.^{30–32} These so-called predictive power procedures use weighted averages of conditional power over values of the alternative, specified through a distribution

$$\Pr(Z(1) \in R | D) = \int \Pr(Z(1) \in R | D, H) \Pr(H | D) dH \quad (14.6)$$

A Bayesian formulation is a natural setting for this approach. If a noninformative prior distribution is used for the distribution of the parameter of interest (e.g., H expressed as a difference of means, difference or ratio of proportions, or hazard ratio), then the posterior distribution is a weighted average of conditional power with equal prior weight for all alternatives. Alternatively, the current (observed) alternative could be used in the conditional power formulation in Equation (14.5) to project power resulting from further follow-up according to the pattern of observations thus far.^{27,31,32} Informative prior distributions that assign higher weight to a specific range

of alternatives may be used in a fully Bayesian approach to assessing interim analysis results, and we describe one such method in the following section.

14.4 A BAYESIAN APPROACH TO ASSESSING EARLY TERMINATION FOR NO BENEFIT

Recently, interest has grown in the application of Bayesian statistical methodology to problems in clinical trial monitoring and early stopping.^{33–37} While Bayesian analyses entail the difficult and sometimes controversial task of specifying prior distributions, if the goal of any clinical trial is ultimately to influence clinical practice, its results must be sufficiently strong to prove compelling to a community of clinical researchers whose prior opinions and experiences are diverse. Thus, in situations where early termination is considered, an analysis of the robustness of conclusions over a range of priors thought to resemble the *a priori* beliefs of reasonable members of the clinical research community should provide insight into the impact that trial results might be expected to exert on clinical practice. This is often an overlooked aspect of trial monitoring, as early stopping can result in diminished impact of the findings and continued controversy and delay while results are debated and large, expensive trials are replicated.

Bayesian calculations for clinical trial monitoring can be motivated by adopting the log hazard ratio δ defined in Section 14.2 as a summary measure of relative treatment efficacy.³⁴ We denote the partial maximum likelihood estimate of the log hazard ratio as $\hat{\delta}$. The parameter δ has an approximately normal likelihood with mean $\hat{\delta}$ and variance $4/m$, where m is the total number of events observed thus far. We assume a normal prior distribution for δ with specified mean δ_p and variance σ_p^2 . The values of δ_p and σ_p^2 may be determined to reflect an individual's prior level of enthusiasm regarding the efficacy of a proposed regimen, and these parameters may be altered to reflect varying degrees of enthusiasm. In this spirit, the notion of "skeptical" and "optimistic" prior distributions is discussed by several authors.^{33–35,38} It is suggested that a skeptical member of the clinical community may adopt a prior for δ that is centered at 0 reflecting the unfortunate fact that relatively few regimens tested lead to material improvements in outcome. Nevertheless, the trial designers will have specified a planning alternative for δ , $\delta = \delta_A$, say, which they must feel is both clinically meaningful and relatively probable. If the skeptic is reasonably open-minded, he or she would be willing to admit some probability that this effect could be achieved, perhaps on the order of 5%. Using these considerations, a skeptical prior with mean $\delta_p = 0$ and standard deviation $\sigma_p = \delta_A/1.645$ is specified. By similar logic, one might be inclined to consider the trial organizers as being among the most optimistic of its proponents, but even they would be reasonably compelled to admit as much as a 5% chance that the proposed regimen will have no effect, i.e., that $\delta \leq 0$. It may therefore be reasonable to model an optimist's prior by setting $\delta_p = \delta_A$ and $\sigma_p = \delta_A/1.645$.

For a given prior distribution and the observed likelihood, an approximate posterior density can be obtained for δ , and the current weight of evidence for benefit can thus be assessed directly by observing the probability that the effect is in some specified range of interest, say $\delta > 0$, indicating a benefit, or $\delta \geq \delta_{ALT} > 0$, corresponding to some clinically relevant effect size δ_{ALT} . Following well-known results from Bayesian

inference using the normal distribution, the posterior distribution for δ has mean and variance given by

$$\delta_{\text{post}} = (n_0\delta_p + m\hat{\delta})/(n_0 + m) \text{ and } \sigma_{\text{post}}^2 = 4/(n_0 + m)$$

where $n_0 = 4/\sigma_p^2$. This quantity is thought of as the prior “sample size” because the information in the prior distribution is equivalent to that in a hypothetical trial yielding a log hazard ratio estimate of δ_p based on this number of events.

From the posterior distribution one can also formulate a predictive distribution in order to assess the consequences of continuing the trial for some fixed additional number of failures. As before, let m be the number of events observed thus far, and let n be the number of additional events to be observed. Denote by δ_n the log relative risk that maximizes that portion of the partial likelihood corresponding to failures $m + 1, m + 2, \dots, m + n$. The predictive distribution of δ_n is normal with the same mean as the posterior distribution and variance $\sigma_{\text{pred}}^2 = 4/(n_0 + m) + 4/n$.

14.5 EXAMPLES

14.5.1 A TRIAL STOPPED EARLY FOR NO BENEFIT

In 1982, NSABP initiated Protocol B-14, a double-blind randomized trial comparing five years of tamoxifen (10 mg b.i.d.) with placebo in patients having estrogen receptor-positive breast tumors and no axillary node involvement. The first report of findings in 1989 indicated improved disease-free survival (DFS, defined as time to either breast cancer recurrence, contralateral breast cancer or other new primary cancer, or death from any cause, 83% vs. 77% event-free at 4 years, $p < 0.00001$). Subsequent follow-up through 10 years has confirmed this benefit, with 69% of patients receiving tamoxifen remaining event-free compared to 57% of placebo patients, and has also shown a significant survival advantage (at 10 years, 80% tamoxifen vs. 76% placebo, $p = 0.02$).³⁹

In April 1987 a second randomization was initiated. Patients who had received tamoxifen and were event-free through five years were rerandomized to either continue tamoxifen for an additional 5 years or to receive placebo. Between April 1987 and December 1993, 1172 patients were rerandomized. To provide for a 0.05 level one-sided test with a power of at least 0.85 under the assumed alternative of a 40% reduction in DFS failure rate, a total of 115 events would be required prior to definitive analysis. Four interim analyses were scheduled at approximately equal increments of information time. Stopping boundaries were obtained using the method of Fleming et al.¹⁰ at the two-sided 0.10 level. Confidential interim end-point analyses were to be compiled by the study statistician and presented to the independent Data Monitoring Committee (DMC) of the NSABP.

At the first interim analysis, based on all data received as of September 30, 1993, more events had occurred in the tamoxifen group (28 events) than among those receiving placebo (18 events). There had been six deaths on the placebo arm and nine among tamoxifen patients. By the second interim analysis (data received as of September 30, 1994), there were 24 events on the placebo arm and 43 on the tamoxifen arm (relative

risk = 0.57, nominal $2p = 0.03$), with ten deaths on the placebo arm and 19 among patients receiving tamoxifen. While there was concern regarding the possibility of a less favorable outcome for patients continuing tamoxifen, we recommended that the trial be continued to the next scheduled interim analysis because the early stopping criterion was not achieved ($2\alpha = 0.0030$) and follow-up for the majority of patients was relatively short (mean 3.75 years). At that time, we computed the conditional probability of rejecting the null hypothesis at the scheduled final analysis (115 events), given the current data and a range of alternative hypotheses (Equation 14.5). Results suggested that even under extremely optimistic assumptions concerning the true state of nature, the null hypothesis could almost certainly not be rejected: even under the assumption of a 67% reduction in failures, the conditional probability of eventual rejection was less than 5%. We also considered an “extended trial” repeating conditional power calculations as if we had intended to observe a total of 229 events before final analysis (this number of events would allow for the detection of a 30% reduction in event rate with a power of 85%). Results indicated that if the trial was continued and the underlying relative risk actually was strongly in favor of tamoxifen, then the null hypothesis possibly could be rejected (Figure 14.1).

At the third interim analysis (data received as of June 30, 1995) there were 32 events on the placebo arm and 56 on the treatment arm (relative risk = 0.59). Four-year DFS was 92% for patients on placebo and 86% for patients on tamoxifen. The

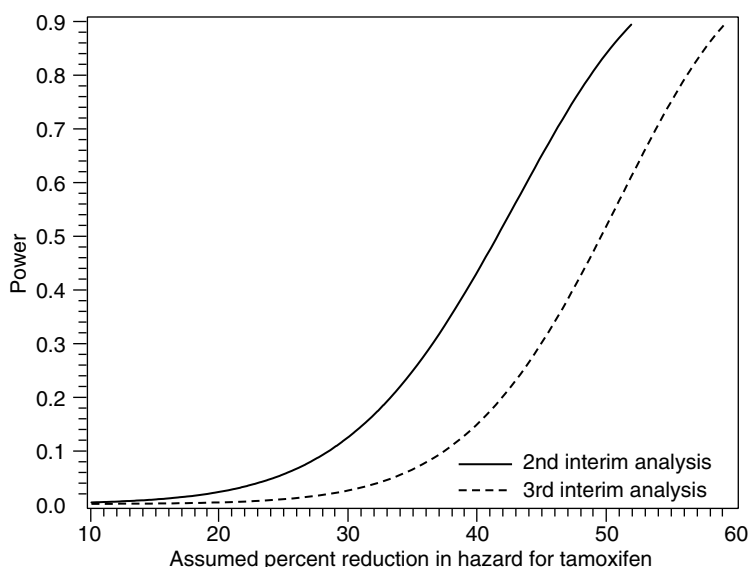


FIGURE 14.1 Conditional probability of finding a significant benefit for tamoxifen in the NSABP B-14 trial if definitive analysis were deferred to the 229th event. Probabilities are conditional on the results of the second and third interim analysis and are graphed as a function of the assumed placebo/tamoxifen relative risk translated to a percent reduction in failure risk for tamoxifen. The curves are calculated using Equation (14.5).²⁸ (Adapted with permission of Elsevier Science from Dignam, J.J. et al., *Control. Clin. Trials*, 19, 575, 1998.)

boundary for early termination ($2\alpha = 0.0035$) was not crossed ($2p = 0.015$). However, calculations showed that even if the remaining 27 events (of 115) all occurred on the placebo arm, the logrank statistic would not approach significance. The imbalance in deaths also persisted (13 placebo arm, 23 tamoxifen, $2p = 0.11$). For the extended trial allowing follow-up to 229 events, Figure 14.1 shows that conditional power was now about 15% under the planning alternative of 40% reduction in relative risk and was only 50% under the more unlikely assumption of a two-fold benefit for continuing tamoxifen. At this time, we also considered the early stopping rule proposed by Wieand et al.²² discussed earlier. To illustrate the consequences of superimposing this rule on the established monitoring boundaries of this trial, suppose the lower boundaries at the third, fourth, and final analyses were replaced with zeros. Then the (upper) level of significance is reduced from 0.0501 to 0.0496, and the power under the alternative of a 40% reduction in event rate is reduced from 0.8613 to 0.8596. By this interim analysis, considerably more events had occurred on the treatment arm than on the control arm. Had such a conservative futility rule been incorporated into the monitoring plan, it would have suggested termination by this time.

As discussed, frequentist approaches were taken in considering the early termination of the B-14 study. We have also applied Bayesian methods for comparative purposes and to attempt to address the broader question of consensus in clinical trials, as the closure of the B-14 study had prompted some criticism from the cancer research community.^{40–42}

Figure 14.2 shows an optimistic prior distribution for the placebo/tamoxifen log hazard ratio, centered at $\delta_p = \delta_A = 0.511$, corresponding to a 40% reduction in failures for patients continuing on tamoxifen relative to those stopping at 5 years. The

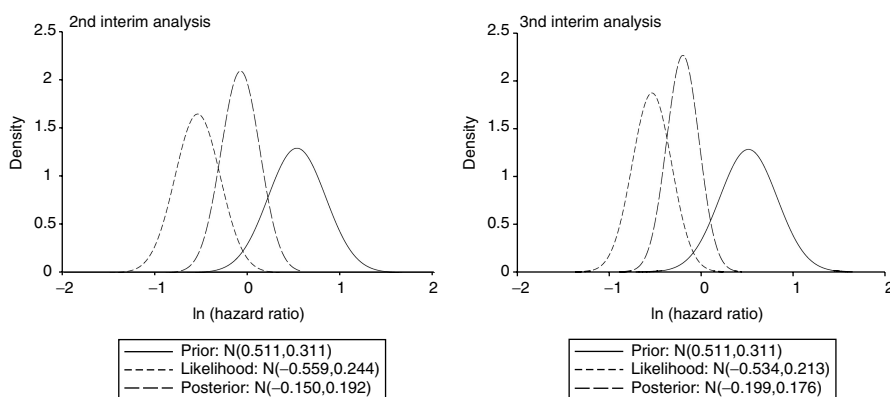


FIGURE 14.2 Prior distribution, likelihood, and posterior distribution of the logged placebo/tamoxifen hazard ratio following the second (left) and third (right) interim analyses of NSABP B-14. An “optimistic” normal prior distribution is assumed, under which the most probable treatment effect is a 40% reduction in failure risk with only a 5% prior probability that the treatment provides no benefit. The resulting posterior distribution at the third interim analysis contains about 13% probability mass to the right of $\ln(\text{hazard ratio}) \delta = 0$. (Adapted with permission of Elsevier Science from Dignam, J.J. et al., *Control. Clin. Trials*, 19, 575, 1998.)

prior standard deviation is $\sigma_p = \delta_A/1.645 = 0.311$. The log hazard ratio likelihood for the B-14 data at the second interim analysis (Figure 14.2, left) has mean $\ln(0.549) = -0.599$ and standard deviation 0.244. The resulting posterior distribution has mean -0.150 and standard deviation 0.192. From this distribution one can determine that the posterior probability that $\delta > 0$ is $1 - \Phi(0.150/0.244) = 0.22$, where $\Phi(\cdot)$ is the standard normal distribution function. By the third interim analysis (Figure 14.2, right), the likelihood has mean $\ln(0.586) = -0.534$ (standard deviation 0.213), resulting in a posterior distribution centered at -0.199 (standard deviation 0.176) and a posterior probability that $\delta > 0$ of 0.13. To the degree that this prior distribution represents that of a clinical researcher who was initially very optimistic, these calculations suggest that even in the face of the negative trial findings as of the third interim analysis, this individual would still assign a small but non-negligible probability to the possibility that continued tamoxifen has some benefit. On the other hand, this individual would now assign essentially no probability ($\approx 3 \times 10^{-5}$) to the possibility that the benefit is as large as a 40% reduction in risk.

For the prior distribution specified above and the observations at the third interim analysis, we also computed the predictive distribution. If the trial were to be extended to allow for a total of 229 events, or 141 events beyond the third interim analysis, we obtain $\delta_{\text{pred}} = -0.199$ and $\sigma_{\text{pred}} = 0.243$. The predictive probability of a treatment comparison achieving nominal significance following the 229th event is determined as follows: if δ_{m+n} denotes the estimated log relative risk based on all the data, then the ultimate result will be significant at the 0.05 level if $\delta_{m+n} > 1.645 \cdot \sqrt{(4/229)} = 0.217$. Since approximately $\delta_{m+n} \approx (88\delta + 141\delta_n)/229$, this circumstance requires that $\delta_n > 0.686$. The predictive probability of this occurrence is $1 - \Phi(\{0.686 + 0.199\}/0.243) \approx 0.0001$.

Subsequent follow-up of Protocol B-14 continues to support the findings that prompted early closure of this study. By one year subsequent to publication (data through December 31, 1996), 135 total events had occurred, 85 among patients who had received continued tamoxifen and 50 among patients rerandomized to placebo (relative risk = 0.60, nominal $2p = 0.002$). There were 36 deaths among tamoxifen patients and 17 among control patients (nominal $2p = 0.01$). A more extensive discussion of this case study has been published elsewhere.⁴³ Published longer term findings from this trial⁴⁴ and one of a similar design by other investigators⁴⁵ remain consistent with the initial conclusions, although because therapy was discontinued, these results become somewhat more difficult to interpret with respect to the original trial questions.

14.5.2 THE EFFECT OF TWO EASILY APPLIED RULES IN THE ADJUVANT AND ADVANCED DISEASE SETTING

The trial presented in the preceding example was not designed with an asymmetric rule for stopping in the face of negative results. It was partly for this reason that the data monitoring committee and investigators considered several methods of analysis before reaching a decision as to whether to stop the trial. Although there will sometimes be special circumstances that require analyses not specified *a priori*, it is preferable to determine in advance whether the considerations for stopping are asymmetric in nature and, if so, to include an appropriate asymmetric stopping rule in the initial trial design.

There are computer packages (EaSt, Cytel Software Corp., Cambridge, MA; and PEST4, MPS Research Unit, University of Reading, Reading, UK) available to help with the design of studies using any of the rules discussed in Section 14.2 (see Emerson⁴⁶ for a review). Alternatively, one may modify a “standard” symmetric rule (e.g., O’Brien–Fleming boundaries) by retaining the upper boundary for early stopping due to positive results but replacing the lower boundary to achieve a more appropriate rule for stopping due to negative results. It is often the case that this will alter the operating characteristics of the original plan so little that no additional iterative computations are required.

To illustrate this, suppose one had designed a trial to test the hypothesis $H_0: \delta = 0$ versus $\delta > 0$ to have 90% power versus the alternative $H_A: \delta = \ln(1.5)$ with a one-sided $\alpha = 0.025$ using a Fleming et al.¹⁰ rule with three interim looks. Such a design would require interim looks when there had been 66, 132, and 198 events with final analysis at 264 events. From Table 1a of Fleming et al.,¹⁰ one such rule would be to stop and conclude that the treatment was beneficial if the standardized logrank statistic Z exceeded 2.81 at the first look, 2.74 at the second look, or 2.67 at the third look. The null hypothesis would be rejected at the end of the trial if Z exceeded 2.02. If a symmetric lower boundary were considered inappropriate, one might choose to replace it by simply testing the alternate hypothesis $H_A: \delta = \ln(1.5)$ vs. $\delta < \ln(1.5)$ at some small significance level (e.g., $\alpha = 0.005$) at each interim look (this suggestion is adapted from the monitoring rules in Southwest Oncology Group Protocol SWOG-8738, an advanced disease lung cancer trial). In the framework of Section 14.2, this rule is asymptotically equivalent to stopping if the standardized Z is < -0.93 at the first look or < -0.25 at the second look or < 0.28 at the third look (a fact that one does not need to know to use the rule because the alternative hypothesis can be tested directly using standard statistical software, e.g., SAS Proc PHREG).⁴⁷ It follows that if the experimental treatment adds no additional benefit to the standard regimen there would be a 0.18 chance of stopping at the first look, a 0.25 chance of stopping at the second look, and a 0.22 chance of stopping at the third look (Table 14.1). Adding this rule does not significantly change the experiment-wise type I error rate ($\alpha = 0.0247$) and would only decrease the power to detect a treatment effect of $\delta_A = \ln(1.5)$ from 0.902 to 0.899.

TABLE 14.1

Probability of Stopping at Each of Three Early Looks Using Two Easily Applied Futility Monitoring Rules

No. of Events	Probability of Stopping if Treatments Are Equivalent		Probability of Stopping under Alternative $\delta = \ln(1.5)$	
	SWOG	WSO	SWOG	WSO
66	0.18	0.0025	0.005	0.000
132	0.25	0.4975	0.004	0.010
198	0.22	0.10	0.003	0.001

Following Wieand et al.,²² an alternative but equally simple way to modify the symmetric Fleming et al. boundaries would be simply to replace the lower Z-critical values with 0s at each interim analysis at or beyond the half-way point. Using this approach, if the experimental arm offered no additional benefit to that of the standard regimen, the probability of stopping at the first look would be very small (0.0025), but the probability of stopping at the second look would be 0.4975 and at the third look would be 0.10 (Table 14.1). Again, no special program is needed to implement this rule, and its use has a negligible effect on the original operating characteristics of the group sequential procedure ($\alpha = 0.00248$, power = 0.900).

The decision of which rule to use will depend on several factors, including the likelihood that patients will still be receiving the treatment at the time of the early looks and the likelihood that the experimental treatment would be used outside the setting of the clinical trial before its results are presented. To illustrate this, we consider two scenarios.

14.5.2.1 Scenario 1

The treatment is being tested in an advanced disease trial where the median survival with conventional therapy has been 6 months and the alternative of interest is to see if the experimental treatment results in at least a 9-month median survival. Under the assumption of constant hazards, this is equivalent to the hypothesis $\delta_A = \ln(1.5)$. Suppose one would expect the accrual rate to such a study to be 150 patients per year. If one designed the study to accrue 326 patients, which would take 26 months, one would need to follow them for slightly less than 4.5 additional months to observe 264 deaths if the experimental regimen offers no additional benefit to the standard regimen or an additional 7.5 months if $\delta = \delta_A = \ln(1.5)$. If the experimental treatment offers no additional benefit, one would expect 146 patients to have been entered when 66 deaths have occurred, 227 patients to have been entered when 132 deaths have occurred, and 299 patients to have been entered when 198 deaths have occurred (Table 14.2). Early stopping after 66 deaths have occurred would prevent 180 patients from being entered to the trial, and stopping after 132 deaths would prevent 99 patients from being entered. Thus, the potential benefit of stopping in the

TABLE 14.2
Effect of Two Easily Applied Rules on Accrual and Reporting Time

No. of Events	No. of Patients Accrued		No. of Patients to Be Accrued		Time Until Final Analysis (mo)	
	Advanced Disease Trial	Adjuvant Treatment Trial	Advanced Disease Trial	Adjuvant Treatment Trial	Advanced Disease Trial	Adjuvant Treatment Trial
66	146	1975	180	625	19	36
132	227	2600	99	0	12	24
198	299	2600	27	0	7	12
264	326	2600	0	0	0	0

face of negative results would be to prevent a substantial number of patients from receiving the apparently ineffective experimental regimen, in addition to allowing early reporting of the results (the savings in time for reporting the results would be approximately 19, 12, and 7 months according to whether the trial stopped at the first, second, or third look, respectively).

14.5.2.2 Scenario 2

The treatment is being tested in an adjuvant trial where the expected hazard rate is 0.0277 deaths/person-year, corresponding to a 5-year survival rate of slightly more than 87%. If one is now looking for an alternative $\delta_A = \ln(1.5)$ (which would roughly correspond to increasing the 5-year survival rate to 91%), and if the accrual rate were approximately 800 patients per year, a reasonable plan would be to accrue 2600 patients, which would take approximately 39 months, and to analyze the data when 264 deaths have occurred, which should occur approximately 66 months after initiation of the trial if the experimental regimen offers no additional benefit to the standard regimen (75 months after initiation if $\delta = \delta_A = \ln\{1.5\}$). With this accrual and event rate, 1975 of the expected 2600 patients will have been entered by the time 66 events have occurred if the experimental regimen offers no additional benefit to the standard regimen (Table 14.2). The second and third looks would occur approximately 3 and 15 months after the termination of accrual, so early stopping following these analyses would have no effect on the number of patients entering the trial, although it could permit early reporting of the results. The savings in time for reporting the results would be approximately 36, 24, and 12 months according to whether the trial stopped at the first, second, or third look, respectively. If there is little likelihood that the therapy will be used in future patients unless it can be shown to be efficacious in the current trial, there may be little advantage to reporting early negative results, and one might choose not to consider early stopping for negative results at any of these looks.

14.6 SUMMARY

Statistical monitoring procedures are used in cancer clinical trials to ensure the early availability of efficacious treatments while at the same time preventing spurious early termination of trials for apparent benefit that may later diminish. Properly designed, these procedures also can provide support for stopping a trial early when results do not appear promising, conserving resources and affording patients the opportunity to pursue other treatment options and avoid regimens that may have known and unknown risks while offering little benefit. By weighing these considerations against each other in the specific study situation at hand, a satisfactory monitoring procedure can be chosen.

The group sequential monitoring rules described in this chapter differ with respect to their operating characteristics, and care should be taken to select a monitoring policy that is consistent with the goals and structure of a particular clinical trial. For example, among symmetric rules, the Pocock procedure is associated with a relatively large maximum number of events (m_K) required for final analysis, but because the probability of early stopping under alternatives of significant treatment

effect is relatively high, the expected number of events required to trigger reporting of results is reduced under such alternatives. In contrast, for the O'Brien–Fleming procedure, m_K is only very slightly greater than the number of events that would be required if no interim analyses were to be performed. The price paid for this is some loss of efficiency (in terms of expected number of events required for final analysis) under alternatives of significant treatment effect. In phase III cancer trials (particularly in the adjuvant setting), it is often the case that both accrual and treatment of patients are completed before a significant number of clinical events (e.g., deaths or treatment failures) have occurred, and more emphasis has been placed on the use of interim analysis policies to prevent the premature disclosure of early results than on their use to improve efficiency by allowing the possibility of early reporting. In such circumstances, it has often been considered to be most important to minimize the maximum number of required events, leading to the rather widespread use of the O'Brien–Fleming method and similar methods such as those of Haybittle and Fleming et al. Other considerations (e.g., the need to perform secondary subset analyses, the possibility of attenuation of treatment effect over time) also argue for the accumulation of a substantial number of events before definitive analysis and therefore favor policies that are rather conservative in terms of early stopping. The power boundaries of Wang and Tsiatis provide a convenient way to explore trade-offs between maximum event size and expected number of events to final analysis by considering a variety of values of the tuning parameter Δ . Emerson, Kittelson, and Gillen²⁹ present a comprehensive approach to the evaluation of operating characteristics for monitoring plans and specifically discuss the assessment of plans that include futility monitoring with respect to important parameters such as expected sample size, probability of early stopping at each analysis, and properties of the treatment effect estimate obtained under the stopping decision.

In the absence of a prespecified stopping rule that is sensitive to the possibility of early stopping for no benefit or negative results, such evidence of a negative effect for an experimental therapy at the time of an interim analysis may be analyzed with the help of conditional power calculations, predictive power, or fully Bayesian methods. These methods are quite practical, have had a careful mathematical development, and have well-studied operating characteristics. We discussed these approaches in Sections 14.3 and 14.4 and applied several of them to data from the NSABP Protocol B-14, a study that was closed in the face of negative results for an experimental schedule of extended tamoxifen (10 years vs. the standard 5 years). In fact, Bayesian monitoring plans based on the use of the skeptical and optimistic prior distributions as described earlier have been used in several cancer clinical trials in the United Kingdom. This approach has been found to aid the trial data monitoring committees in focusing on the treatment effect size and whether interim results had effectively established or ruled out the possibility of a clinically relevant benefit (rather than significance levels for a given alternative) for the therapy in question, which the authors believe eventually contributed to more widely convincing trial findings.⁴⁸

It is certainly preferable to include a plan for stopping in the face of negative results at the time the study protocol is developed. In particular, it is important to know what effect the plan will have on the power and significance level of the overall design. The mathematics required to create an appropriate group sequential design that

incorporates asymmetric monitoring boundaries is presented in Section 14.2, with examples of several asymmetric designs in current use. Many factors enter into the choice of a design including the anticipated morbidity of the experimental regimen, severity of the disease being studied, the expected accrual rate of the study, and the likely effect of early release of results on other studies. The example in Section 14.5.2 showed that a rule applied in an advanced disease setting might prevent the accrual of a fraction of patients to an ineffective experimental regimen while the same rule applied to an adjuvant trial is likely only to affect the timing of the presentation of results, as accrual may be completed before the rule is applied. When asymmetrical monitoring boundaries are required, our preference has been to use simple approaches, such as the Wieand et al. modification of symmetric boundaries, or to use an upper boundary of the O'Brien–Fleming or Fleming et al. type coupled with a lower boundary derived by testing the alternative hypothesis $H_A: \delta = \delta_A$, using the O'Brien–Fleming or Fleming et al. rules. In the latter case, one may or may not require the procedure to be closed (that is, require the upper and lower boundaries to join at the K th analysis). If closure is required, the use of O'Brien–Fleming boundaries leads to the method of Pampallona and Tsiatis²¹ with $\Delta = 0$. Our experience is that these approaches are easily explained to (and accepted by) our clinical colleagues.

We advocate that trial designers give serious thought to the suitability of asymmetric monitoring rules. If such an approach is reasonable, the methods in Section 14.2 allow the statistician to develop a design that seems most appropriate for his or her situation. If at an interim look, one is faced with negative results and has not designed the trial to consider this possibility, we recommend one of the approaches described in Sections 14.3 and 14.4. Even when one has had the foresight to include an asymmetric design, one may gain further insight regarding unexpected results by applying some of these methods. Of course, when one deviates from the original design, the initial power and significance level computations of the study are altered.

Combined with these recommendations, some caution is also warranted when specifying a strategy for futility monitoring, given the effort and cost required in mounting a phase III clinical trial. Friedlin and Korn⁴⁹ recently identified and discussed four major concerns with aggressive monitoring for lack of activity: 1) some rules may dictate stopping for futility when in fact the experimental arm is superior to the control arm at a given interim analysis; 2) certain asymmetric boundaries may call for early stopping when the conditional power criteria implied by such bounds are not particularly small, say $\gamma_A = 0.50$; 3) with futility stopping, there is a consequent loss of power to detect treatment differences that are smaller than those of the trial design specifications but nonetheless of clinical relevance; and 4) there is sensitivity to departures from the proportional hazards assumption of power and any projections about eventual findings had the trial continued. They evaluated an aggressive futility rule using the approach of Pampallona and Tsiatis²¹ and two more conservative rules, a stochastic curtailment approach with conditional stopping probability $\gamma_A = 0.10$ and a rule akin to that used by SWOG described in Fleming et al.¹⁰ Among these concerns, aggressive early stopping rules were most problematic with respect to potentially inappropriate stopping in cases where small positive treatment effects (e.g., those less than H_A) were present in interim data. This is a realistic concern because trials are often planned with overly favorable alternatives rather than the

minimum clinically relevant treatment effect. Their findings did not show a large effect of nonproportionality, but obviously if the constant hazard ratio assumption were seriously in error, then any projections from interim data would be flawed. However, it may be more likely in cancer trials for an early difference to later vanish than for a late difference to emerge.

This chapter provides only a brief sketch of this important area in oncology clinical trials, and we recommend the comprehensive texts by Whitehead¹⁶ and Jennison and Turnbull.⁵⁰ An advance copy of the latter was kindly provided to us in 1999 as we completed the first edition version of this chapter.

REFERENCES

1. Tsiatis, A.A. The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika*, 68, 311, 1981.
2. Tsiatis, A.A. Repeated significance testing for a general class of statistics used in censored survival analysis. *J. Am. Stat. Assoc.*, 77, 855, 1982.
3. Gail, M.H., DeMets, D.L., and Slud, E.V. Simulation studies on increments of the two-sample logrank score test for survival time data, with application to group sequential boundaries. In *Survival Analysis*, Crowley, J. and Johnson, R.A., eds., Hayward, CA: Institute of Mathematical Statistics Lecture Notes-Monograph Series, vol. 2, 1982, 287–301.
4. Armitage, P., McPherson, C.K., and Rowe, B.C. Repeated significance tests on accumulating data. *J. Royal Stat. Soc. Series A*, 132, 235, 1969.
5. Haybittle, J.L. Repeated assessment of results in clinical trials in cancer treatments. *Br. J. Radiol.*, 44, 793, 1971.
6. Pocock, S.J. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64, 191, 1977.
7. Pocock, S.J. Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics*, 38, 153, 1982.
8. O'Brien, P.C. and Fleming, T.R. A multiple testing procedure for clinical trials. *Biometrics*, 35, 549, 1970.
9. Wang, S.K. and Tsiatis, A.A. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 43, 193, 1987.
10. Fleming, T.R., Harrington, D.P., and O'Brien, P.C. Designs for group sequential tests. *Control. Clin. Trials*, 5, 348, 1984.
11. DeMets, D.L. and Ware, J.H. Group sequential methods for clinical trials with a one-sided hypothesis. *Biometrika*, 67, 651, 1980.
12. Wald, A. *Sequential Analysis*. New York: John Wiley and Sons, 1947.
13. DeMets, D.L. and Ware, J.H. Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika*, 69, 661, 1982.
14. Whitehead, J. and Stratton, I. Group sequential clinical trials with triangular continuation regions. *Biometrics*, 39, 227, 1983.
15. Whitehead, J. and Jones, D. The analysis of sequential clinical trials. *Biometrika*, 66, 443, 1979.
16. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. Rev. 2nd ed. Chichester: Wiley 1997.
17. Whitehead, J. and Matsushita, T. Stopping clinical trials because of treatment ineffectiveness: a comparison of a futility design with a method of stochastic curtailment. *Stat. Med.*, 22, 677, 2003.

18. Jennison, C. Efficient group sequential tests with unpredictable group sizes. *Biometrika*, 74, 155, 1987.
19. Lan, K.K.G. and DeMets, D.L. Discrete sequential boundaries for clinical trials. *Biometrika*, 70, 659, 1983.
20. Emerson, S.S. and Fleming, T.R. Symmetric group sequential test designs. *Biometrics*, 45, 905, 1989.
21. Pampallona, S. and Tsiatis, A.A. Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *J. Stat. Plan. Infer.*, 42, 19, 1994.
22. Wieand, S., Schroeder, G., and O'Fallon, J.R. Stopping when the experimental regimen does not appear to help. *Stat. Med.*, 13, 1453, 1994.
23. Ellenberg, S.S. and Eisenberger, M.A. An efficient design for phase III studies of combination chemotherapies. *Cancer Treatment Rep.*, 69, 1147, 1985.
24. Wieand, S. and Therneau, T. A two-stage design for randomized trials with binary outcomes. *Control. Clin. Trials*, 8, 20, 1987.
25. Lan, K.K.G., Simon, R., and Halperin, M. Stochastically curtailed tests in long-term clinical trials. *Commun. Stat. — Sequent. Anal.*, 1, 207, 1982.
26. Halperin, M. et al. An aid to data monitoring in long-term clinical trials. *Control. Clin. Trials*, 3, 311, 1982.
27. Lan, K.K.G. and Wittes, J. The B-value: a tool for monitoring data. *Biometrics*, 44, 579, 1988.
28. Davis, B.R. and Hardy, R.J. Upper bounds for type I and type II error rates in conditional power calculations. *Commun. Stat. Theory Meth.*, 19, 3571, 1991.
29. Emerson, S.S., Kittelson, J.M., and Gillen, D.L. Evaluation of Group Sequential Clinical Trial Designs (November 7, 2003). UW Biostatistics Working Paper Series. Working Paper 216. <http://www.bepress.com/uwbiostat/paper216>.
30. Spiegelhalter, D.J., Freedman, L.S., and Blackburn, P.R. Monitoring clinical trials: conditional or predictive power? *Control. Clin. Trials*, 7, 8, 1986.
31. Choi, S.C. and Pepple, P.A. Monitoring clinical trials based on predictive probability of significance. *Biometrics*, 45, 317, 1989.
32. Jennison, C. and Turnbull, B.W. Statistical approaches to interim monitoring of medical trials: a review and commentary. *Stat. Sci.*, 5, 299, 1990.
33. Kass, R. and Greenhouse, J. Comment on "Investigating therapies of potentially great benefit: ECMO" by J. Ware. *Stat. Sci.*, 4, 310, 1989.
34. Spiegelhalter, D.J., Freedman, L.S., and Parmar, M.K.B. Bayesian approaches to randomized trials (with discussion). *J. Royal Stat. Soc. Series A*, 157, 357, 1994.
35. Freedman, L.S., Spiegelhalter, D.J., and Parmar, M.K.B. The what, why, and how of Bayesian clinical trials monitoring. *Stat. Med.*, 13, 1371, 1994.
36. Greenhouse, J. and Wasserman, L. Robust Bayesian methods for monitoring clinical trials. *Stat. Med.*, 14, 1379, 1994.
37. Berry D.A., Stangl D.K., eds. *Bayesian Biostatistics*. New York: Marcel Dekker, 1996.
38. Parmar, M.K.B., Ungerleider, R.S., and Simon, R. Assessing whether to perform a confirmatory randomized clinical trial. *J. Natl. Cancer Inst.*, 88, 1645, 1996.
39. Fisher, B. et al. Five versus more than five years of tamoxifen therapy for breast cancer patients with negative lymph nodes and estrogen receptor-positive tumors. *J. Natl. Cancer Inst.*, 88, 1529, 1996.
40. Swain, S.M. Tamoxifen: the long and short of it. *J. Natl. Cancer Inst.*, 88, 1510, 1996.
41. Peto, R. Five years of tamoxifen — or more? *J. Natl. Cancer Inst.*, 88, 1791, 1996.

42. Current Trials Working Party of the Cancer Research Campaign Breast Cancer Trials Group. Preliminary results from the Cancer Research Campaign trial evaluating tamoxifen duration in women aged fifty years or older with breast cancer. *J. Natl. Cancer Inst.*, 88, 1834, 1996.
43. Dignam, J.J. et al. Early stopping of a clinical trial when there is evidence of no treatment benefit: Protocol B-14 of the National Surgical Adjuvant Breast and Bowel Project. *Control. Clin. Trials*, 19, 575, 1998.
44. Fisher, B. et al. Five versus more than five years of tamoxifen for lymph node negative breast cancer patients: updated findings from the National Surgical Adjuvant Breast and Bowel Project B-14 randomized trial. *J. Natl. Cancer Inst.*, 93, 684, 2001.
45. Stewart, H.J., Prescott, R.J., and Forrest, A.P. Scottish adjuvant tamoxifen trial: a randomized study updated to 15 years. *J. Natl. Cancer Inst.*, 93, 456, 2001.
46. Emerson, S.S. Statistical packages for group sequential methods. *Am. Stat.*, 50, 183, 1996.
47. SAS Technical Report P-217, *SAS/STAT Software: the PHREG Procedure*, Version 6, Cary, NC: SAS Institute Inc., 1991. 63 pp.
48. Parmar, M.K.B. et al. Monitoring of large randomized clinical trials: a new approach with Bayesian methods. *Lancet*, 358, 375, 2001.
49. Freidlin, B. and Korn, E.L. A comment on futility monitoring. *Control. Clin. Trials*, 23, 355, 2002.
50. Jennison, C. and Turnbull, B.W. *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC Press, 2000.

15 Design and Analysis of Quality of Life Data

*Andrea B. Troxel, Sc.D., and
Carol M. Moinpour, Ph.D.*

CONTENTS

15.1	Introduction	248
15.1.1	What is QOL and How is it Measured?.....	248
15.1.2	Chapter Organization	249
15.2	Design Issues	250
15.2.1	Timing of Assessments	250
15.2.2	Clinical Significance and Power	250
15.2.3	Composite Variables.....	253
15.3	Types of Missing Data	253
15.3.1	Graphical Description of Missing Data Mechanisms	254
15.3.2	Comparing Missing Data Mechanisms	256
15.4	Longitudinal Analysis Options for “Complete” Data Sets	256
15.4.1	Normal Linear Mixed Model.....	257
15.4.2	Generalized Linear Models	257
15.4.3	Generalized Estimating Equations.....	258
15.4.4	Change-Score Analysis	258
15.4.5	Time-to-Event Analysis	258
15.5	Methods for Analysis with Missing Data	258
15.5.1	Joint Modeling of Measurement and Missingness Processes	258
15.5.2	Weighted GEEs	260
15.5.3	Sensitivity Analysis	261
15.6	Avoiding Pitfalls: Some Commonly Used Solutions.....	261
15.6.1	Substitution Methods	261
15.6.2	Multiple Imputation	262
15.6.3	Adjusted Survival Analyses	262
15.6.3.1	Quality Adjusted Life Years	263
15.6.3.2	Q-TWiST	263
	References	263

15.1 INTRODUCTION

The objective of this chapter is to inform those engaged in clinical trials research about the special statistical issues involved in analyzing longitudinal quality of life (QOL) outcome data. Because these data are collected in trials for patients with a life-threatening disease, it is common to see a drop-off in submission rates for patient-reported questionnaire data, often due to death and deteriorating health. Data sets with such missing data must be analyzed appropriately in order not to arrive at misguided conclusions regarding change in QOL over the treatment course. Methods will be suggested for complete or near-complete datasets as well as for those where missing data are a concern.

15.1.1 WHAT IS QOL AND HOW IS IT MEASURED?

In randomized treatment trials for cancer or other chronic diseases, the primary reason for assessing QOL is to broaden the scope of treatment evaluation. We sometimes characterize QOL and cost outcomes as alternative or complementary because they add to information provided by traditional clinical trials' endpoints, such as survival, disease-free survival, tumor response, and toxicity. The challenge lies in combining this information in the treatment evaluation context. There is fairly strong consensus that at least in the phase III setting QOL should be measured comprehensively.¹⁻³ Although a total, or summary, score is desirable for the QOL measure, it is equally important to have separate measures of basic domains of functioning, such as physical, emotional, social, and role functioning, as well as symptom status. Symptoms specific to the cancer site and/or the treatments under evaluation are also usually included to monitor for toxicities and to gauge the palliative effect of the treatment on disease-related symptoms. In some trials, investigators may study additional areas such as financial concerns, spirituality, family well being, and satisfaction with care. Specific components of QOL not only provide information on specific interpretation of treatment effects but also can identify areas in which cancer survivors need assistance in their return to daily functioning. Data on specific areas of functioning can also help suggest ways to improve cancer treatments; Sugarbaker, et al.⁴ conducted a study in which the radiotherapy regimen was modified as a result of QOL data.

QOL data should be generated by patients in a systematic, standardized fashion. Interviews can be used to obtain these data, but self-administered questionnaires are usually more practical in the multi-institution setting of clinical trials. Selected questionnaires must be reliable and valid⁵ and sensitive to change over time;^{6,7} good measurement properties, along with appropriate item content, help ensure a more accurate picture of the patient's QOL. There are four QOL questionnaires that meet these measurement criteria, provide comprehensive measures of domains or areas of QOL, and are frequently used in cancer clinical trials. The Functional Assessment of Chronic Illness Therapy (FACIT), the newer name for the Functional Assessment of Cancer Therapy (FACT, version 4),⁸⁻¹² and the European Organization of Research and Therapy for Cancer (EORTC) Quality of Life Questionnaire — Core 30 (QLQ-C30)¹³⁻¹⁷ are two QOL questionnaires that measure general domains

or areas of QOL, the core component, along with symptom modules specific to the disease or type of treatment; see Web sites for currently available modules.^{18,19} Others, like the Cancer Rehabilitation Evaluation System — Short Form (CARES-SF)^{20–23} and the Short Form Health Survey (SF-36, SF-12, SF-8),^{24–35} can be used with any cancer site but may require supplementation with a separate symptom measure to address concerns about prominent symptoms and side effects. The large normative database available for the SF-36 is very useful in interpreting score changes and differences; see the Web site for information on accessing this measure and its databases.³⁶ This questionnaire has been used successfully in cancer clinical trials to detect differences in treatment regimens.³⁷

15.1.2 CHAPTER ORGANIZATION

The inclusion of QOL endpoints in clinical trials must be treated as seriously as any clinical outcome. Most design issues important for obtaining a credible evaluation of clinical endpoints are also important for QOL outcomes. For example, optimal timing of assessments may vary according to the treatment schedule, disease site, and other factors. Issues of clinical significance, which in turn inform sample size and power considerations, are the same, but the databases used to determine clinical significance are not as familiar to clinicians and are still being developed by QOL researchers. The use of composite variables, gaining popularity but associated with difficulties in interpretation, has also been suggested for QOL outcomes. These subjects are discussed in more detail in Section 15.2.

When QOL research is conducted in many and often widely differing institutions, quality control is critical to ensure clean, complete data. The first step is to make completion of a baseline QOL assessment a trial eligibility criterion. Enforcement of the same requirements for both clinical and QOL follow-up data communicates the importance of the QOL data for the trial. Even with the best quality control procedures, submission rates for follow-up QOL questionnaires can be less than desirable, particularly in the advanced stage disease setting. It is precisely in the treatment of advanced disease, however, that QOL data often provide key information about the extent of palliation achieved by an experimental treatment. While this is a rich source of information, data analysis often is complicated by problems of missing information. Patients sometimes fail to complete QOL assessments because they experience negative events, such as treatment toxicities, disease progression, or death. Because not all patients are subject to these missing observations at the same rate, especially when treatment failure or survival rates differ between arms, the set of complete observations is not always representative of the total group; analyses using only complete observations are therefore potentially biased.

Several methods have been developed to address this problem. They range in emphasis from the data collection stage, where attention focuses on obtaining the missing values, to the analysis stage, where the goal is adjustment to properly account for the missing values. In Section 15.3, we first describe different types of missing data. Section 15.4 describes methods that are appropriate for complete or nearly complete data. Section 15.5 presents several methods that have been used to

address incomplete datasets and informative missing data, including sensitivity analyses. Section 15.6 examines approaches for substituting scores for missing data, imputation, and methods for combining survival and QOL outcomes.

15.2 DESIGN ISSUES

15.2.1 TIMING OF ASSESSMENTS

When adding QOL to a cancer clinical trial, it is important to think carefully about possible assessment times. [Table 15.1](#) summarizes some of the factors that should be considered in order to select clinically meaningful time points as well as time points that are “fair” for all treatment arms under study. Most factors require discussion with clinicians involved in the trial. For example, one might want to document QOL status at known points of remission or deterioration for the particular cancer site or to assess patient QOL at the earliest point when an agent could be expected to have a positive effect on the disease to see if there would be a comparable effect on QOL. The clinicians involved in the trial have likely had previous experience with the agent and can be a good source of suggestions for meaningful time points. The table notes factors affecting compliance with the QOL assessment schedule and quality control procedures. Specifying acceptable windows for the QOL assessment times in the protocol helps with quality control procedures downstream. Finally, one can base QOL assessments on events, such as the beginning of new treatment cycles, as opposed to the number of days from randomization. This decision can have implications for systems used to monitor timely submission of questionnaires. Information about delays in treatment administration often does not reach the data center in time to revise expected due dates for questionnaires; in addition, some systems may not be able to accommodate ongoing demands for due date revisions.

15.2.2 CLINICAL SIGNIFICANCE AND POWER

There is increasing emphasis on the importance of addressing the clinical significance of QOL scores in cancer clinical trials and of helping clinicians and clinical trialists interpret what scores and changes in scores mean. A series of seven papers published in the Mayo Clinic Proceedings in 2002 addressed these subjects.^{40–46} A protocol that includes a QOL outcome should stipulate the number of points that reflects a clinically significant difference or change for that questionnaire. Fortunately, this task is feasible using the large databases that exist for some of the more frequently used questionnaires. For example, the Functional Assessment of Cancer Therapy — Lung Cancer (FACT-L) has a separate score for the trial outcome index (TOI), a measure of physical and functional well being and symptoms/concerns. A change of 5–7 points for the TOI has been associated with differences in other clinical measures such as performance status as well as with the amount of change over time in patients with better versus worse prognosis.⁴⁷ If physical and symptom status are of primary interest in the trial, then a measure such as the TOI might be designated the primary QOL outcome variable and sample size determined based on the 5–7 point improvement for one arm versus the other. Given a trial for

TABLE 15.1
Important Variables in the Determination of QOL Assessment Schedules*

Variable	Example/Rationale
Mandatory baseline assessment	Cannot measure change without an assessment prior to the initiation of treatment.
Data collection prior to administration of treatment and/or discussions with clinical staff	Compare patient experience with different regimens after recovery from previous cycle. Avoid biasing patient report based on feedback from medical staff.
Similarity of timing of QOL assessments for all treatment arms	Comparable assessment times for arms problematic when regimens have different administration schedules (e.g., 3-week vs. 4-week cycles). Assessment time can be based on time (e.g., every two weeks from randomization/registration) or on event (e.g., every two treatment cycles).
Natural course of the disease	Known points of remission and deterioration.
Disease stage	1. Early stage disease: Longer follow-up to address survivorship issues, monitor late effects (both positive and negative), and see if patients are able to return to “normal” activities. 2. Late stage disease: Shorter follow-up period because of the potential for missing data. Median survival one basis for length of QOL follow-up.
Effects associated with the treatment course or administration	1. Documentation of acute, short-term side effects or cumulative side effects such as at the end of radiotherapy (XRT). 2. Minimum number of cycles required to see an effect of treatment on QOL.
Timing of important clinical events or monitoring	1. Can assess QOL when patients go off treatment, such as at progression, but results in patient-specific measurement times based on event timing and the possibility of no data for patients who do not experience the event. 2. Pair QOL assessments with clinical monitoring (e.g., tumor measurements) to enhance forms compliance.
Completion of treatment and/or a short time after completion of treatment	E.g., Resolution of mucositis may require 2–4 weeks post-completion of XRT. Treatment arms might be compared at the end of XRT and 2–4 weeks later to see how much better/sooner palliation occurs.
Scheduling issues for special populations Example: End-of-life care	4 factors $\Rightarrow$ suggested weekly assessment schedule (38): 1. length of survival (~ 30 days for terminal patients) 2. variability in deterioration (more pronounced 1–3 weeks prior to death) 3. length of time required to observe effect of intervention 4. burden issues
Compliance with assessment schedule	1. Respondent burden: too many assessments are burdensome and affect the patient’s adherence to the QOL assessment schedule. 2. Institution staff burden can also affect compliance.

(continued)

TABLE 15.1 (Continued)

Variable	Example/Rationale
	3. Data collection and management resources: frequent assessments require more personnel and data management effort and can compromise quality control procedures.
	4. Specification of acceptable time windows: even with specified times of assessment, variability occurs in completion dates. This can affect how QOL is rated and possibly the interpretation of treatment arm comparisons. (39)

*Adapted with permission from Sprangers MAG, Moinpour CM, Moynihan TJ, Patrick DL, Revicki DA, and the Clinical Significance Consensus Meeting Group. *Mayo Clin. Proc.* 2002;77:561–571

advanced stage disease patients, investigators might be interested in seeing if patients in one arm deteriorated by a clinically important amount such as 5–7 FACT TOI points while the other arm remained relatively stable. Clinically important change for the EORTC QLQ-C30 varies with the subscale of interest and other variables such as stage of disease;⁴⁸ Osoba et al.⁴⁹ suggest that a change of 10–20 QLQ-C30 points on a 0–100 scale reflects a moderate effect.

If change for a particular questionnaire has not been documented, it is possible to use effect size as a basis for determining a clinically important difference. Effect size⁵⁰ is a measure of the change or difference in scores relative to variability in the scores. It is calculated in a number of different ways but can be summarized as follows: δ (effect size) = (mean $[\mu]_{\text{Arm 2}}$ – mean $[\mu]_{\text{Arm 1}}$)/standard deviation $[\sigma]_{\text{Arm 1}}$. In this case, one has access to means and standard deviations for this questionnaire. The denominator can also be the standard deviation of the control arm, of a stable group, or of the difference, to name a few variations. See Sprangers et al.⁴⁰ for a summary of these formulas. Cohen⁵⁰ has described effect sizes of 0.2 as small, 0.5 as moderate, and 0.8 as large. If the number of points reflecting clinically significant change is not available, sample sizes can be estimated from standard software using effect size parameters: $\mu_1 = 0$; μ_2 = the minimum δ of interest; $\sigma = 1$; $\alpha = 0.05$; and power = 0.80. One half of a standard deviation can also be used as a benchmark for a moderate-sized effect or change.^{50,51} Sloan et al. noted that a difference of 10 on a 0–100 scale generally reflects a moderate effect.⁵¹

If there are normative data for a questionnaire, one can characterize the meaning of a score by comparing the difference in points observed in the trial to differences reported for other groups of patients. Moinpour et al.³⁷ reported that a median difference of 8 points was found for the SF-36 Mental Health Index²⁵ for men receiving orchiectomy plus the antiandrogen flutamide with a median score of 76 compared to the arm treated with orchiectomy alone with a median score of 84, where a lower score reflects worse emotional well-being. Normative data for the Mental Health Inventory (MHI) indicate that men with benign prostatic hyperplasia and hypertension had a median score of 88, while men with congestive heart failure had a median score of 80.²⁸ The severity of these two medical conditions clearly

differs, with men with congestive heart failure reporting more compromised emotional well-being. These data help establish the clinical meaningfulness of an 8-point difference. Even a 3-point difference on the MHI is clinically important because it is associated with decreases in scores reported by individuals who have just lost their jobs.²⁸ It is also possible to include a single item rating of change in QOL and use this anchor to determine the number of points required for clinical significance.⁴²

15.2.3 COMPOSITE VARIABLES

Another issue that arises in cancer clinical trials with multiple traditional clinical endpoints is the advisability of creating composite variables to examine effectiveness; these composite variables can include outcomes such as hospital days, patient-reported symptoms, or general QOL. For example, Freemantle, et al.⁵² specified a composite variable including all-cause mortality, nonfatal myocardial infarction, and refractory ischemia. They reported that the composite variable in the four trials favored treatment with the inhibitor; however, all-cause mortality, a critical component of the composite variable, did not favor the new treatment. This shows the importance of describing effects for all components of a composite variable so that a composite variable does not mask negative or null effects of a more clinically salient outcome. Johnson et al.⁵³ noted that all components of a composite variable need to be related and should have similar clinical importance in order to reflect clinical benefit. Composite variables can increase the statistical precision and efficiency of a trial when the individual components behave in the same direction. However, the results are not clinically helpful when components differ in the size and direction of the treatment effect. Freemantle et al.⁵² suggested that interpretation of results for a composite variable should make clear that these results apply to the combination and not necessarily to individual components. The authors also note that the components should be treated as secondary outcomes with results reported separately for each.

15.3 TYPES OF MISSING DATA

As mentioned briefly above, QOL data are often subject to missingness. Depending on the nature of the mechanism producing the missing data, analyses must be adjusted differently. Below, we list three types of missing data and provide general descriptions of the mechanisms along with their more formal technical names and terms.

The least problematic type of missing data is *missing completely at random* data (MCAR); this mechanism is sometimes termed sporadic. Missing data probabilities are independent of both observable and unobservable quantities; observed data are a random subsample of complete data. This type of mechanism rarely obtains in real data.

Data are said to be *missing at random* (MAR) when the missing data probabilities are dependent on observable quantities, such as previously measured QOL outcomes or covariates like age, sex, and stage of disease, and the analysis can generally be adjusted by weighting schemes or stratification. This type of mechanism can hold

if subjects with poor baseline QOL scores are more prone to missing values later in the trial or if an external measure of health, such as the Karnofsky performance status, completely explains the propensity to be missing. Because the missingness mechanism depends on observed data, analyses can be conducted that adjust properly for the missing observations.

The most difficult type of missing data to handle is termed *nonrandom, missing not at random* (MNAR), or *nonignorable* (NI). Here, missing data probabilities are dependent on unobservable quantities, such as missing outcome values or unobserved latent variables describing outcomes such as general health and well being. This type of mechanism is fairly common in QOL research. One example is treatment-based differences in QOL compliance due to worse survival on one arm of the trial. Another is when subjects having great difficulty coping with disease and treatment are more likely to refuse to complete a QOL assessment.

To determine which methods of statistical analysis will be appropriate, the analyst must first determine the patterns and amount of missing data and identify potential mechanisms that could have generated missing data. In general, there is no reliable way to test for a given type of missing data mechanism, and thus sensitivity analyses are a crucial component of any analysis; this is discussed in more detail below. Rubin⁵⁴ addressed the assumptions necessary to justify ignoring the missing data mechanism and established that the extent of ignorability depends on the inferential framework and the research question of interest. Under likelihood-based and Bayesian inference, the missing data are said to be *ignorable* if the missing data mechanism is MAR and the parameters of the missing data model are distinct from those of the model of interest for the outcome. Identification of missing data mechanisms in QOL research proceeds through two complementary avenues: (1) collecting as much additional patient information as possible and applying simple graphical techniques, and (2) using hypothesis testing to distinguish missing data processes subject to modeling assumptions about the missingness mechanism.

15.3.1 GRAPHICAL DESCRIPTION OF MISSING DATA MECHANISMS

Graphical presentations can be crucial as a first step in elucidating the relationship of missing data to the outcome of interest and providing an overall summary of results that is easily understood by nonstatisticians. A clear picture of the extent of missing QOL assessments is necessary both for selection of the appropriate methods of analysis and for honest reporting of the trial with respect to reliability and generalizability. In clinical trials, this means summarizing the proportions of patients in whom assessment is possible, such as surviving patients still on study, and then the pattern of assessments among these patients. Machin and Weeden⁵⁵ combine these two concepts in [Figure 15.1](#) using the familiar Kaplan–Meier plot to indicate survival rates and a simple table describing QOL assessment compliance. For this study of palliative treatment for patients with small cell lung cancer (SCLC) and poor prognosis, the Kaplan–Meier plot illustrates why the expected number of assessments is reduced by 60% at the time of the final assessment. The table further indicates the increase in missing data even among surviving subjects, from 25% at baseline to 71% among the evaluable patients at six months. If the reasons for missing assessments differ over

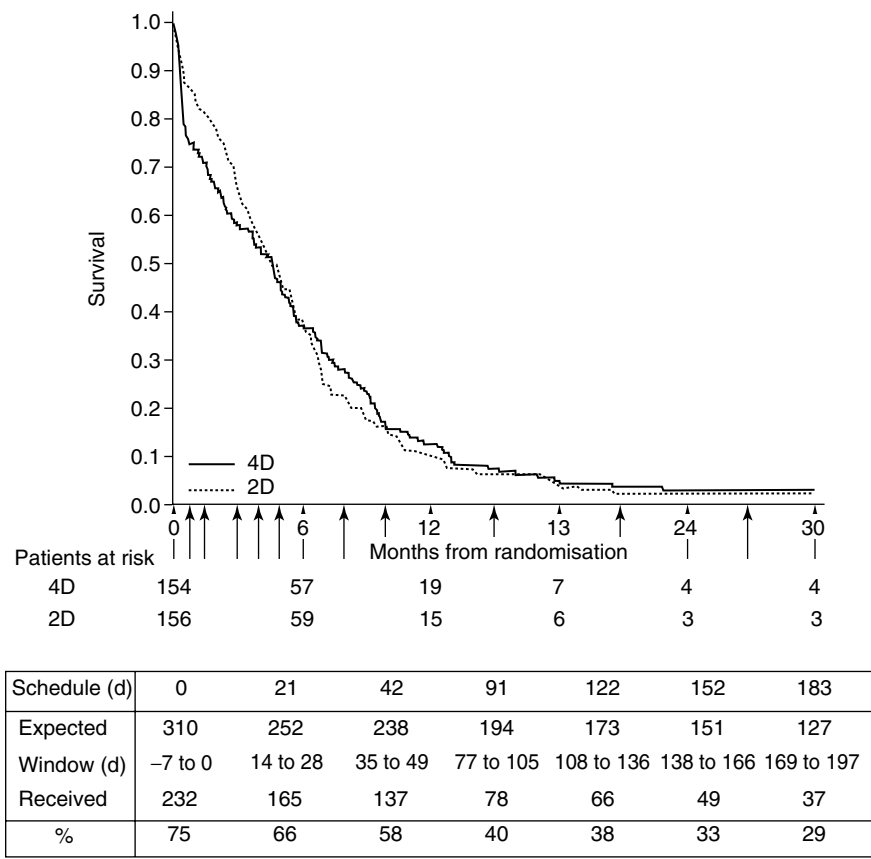


FIGURE 15.1 Kaplan–Meier estimates of the survival curves of patients with SCLC by treatment group (after MRC Lung Cancer Working Party, 1996). The times at which QOL assessments were scheduled are indicated beneath the time axis. The panel indicates the QOL assessments made for the seven scheduled during the first 6 months as a percentage of those anticipated from the currently living patients. (Reprinted with permission from Machin D and Weeden S. *Stat Med* 1998;17:711–724, copyright John Wiley & Sons Limited.)

time or across treatment groups, it may be necessary to present additional details about the missing data.

A second step is describing the missing data mechanism, especially in relation to the patients’ QOL. A useful technique is to present the available data separately for patients with different amounts of and reasons for dropout. This is illustrated by Figure 15.2, due to Troxel,⁵⁶ where estimates of average symptom distress in patients with advanced colorectal cancer are presented by reason for dropout and duration of follow-up. Patients who drop out due to death or illness report higher symptom distress, and the worsening of symptom status over time is more severe for these patients as well. Patients with a decreasing QOL score may also be more likely to drop out, as demonstrated by Curran, et al.,⁵⁷ where a change score between two previous assessments was predictive of dropout.

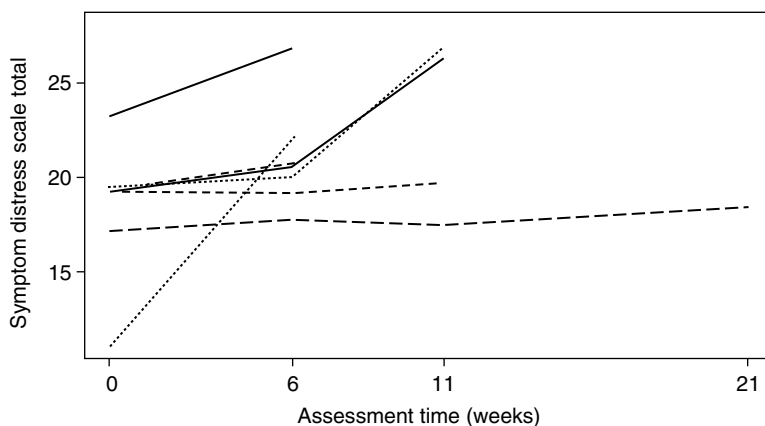


FIGURE 15.2 Average Symptom Distress Scale scores by type and length of follow-up: — lost-death; ···· lost-illness; --- lost-other; ——— complete follow-up. (Reprinted with permission from Troxel AB, *Stat Med* 1998; 17:767, copyright John Wiley & Sons Limited.)

15.3.2 COMPARING MISSING DATA MECHANISMS

Assuming a monotone pattern of missing data, Diggle⁵⁸ and Ridout⁵⁹ have proposed methods to compare MCAR and MAR dropout. The former proposal involves testing whether scores from patients who drop out immediately after a given time point are a random sample of scores from all available patients at that assessment. The latter proposal centers on logistic regression analysis to test whether observed covariates affect the probability of dropout.

Testing the assumptions of MAR against a hypothesis of MNAR is not trivial; such a procedure rests on strong assumptions that are themselves untestable.⁵⁷ When fitting a NI model, certain assumptions are made in the specification of the model about the relationship between the missing data process and unobserved data. These assumptions are fundamentally untestable. Molenberghs et al.⁶⁰ provide examples where different models produce almost similar fits to the observed data but yield completely different predictions for the unobserved data. Little,⁶¹ discussing pattern-mixture models, suggests that underidentifiability is a serious problem with MNAR missing data models and that problems may arise when estimating the parameters of the missing data mechanism simultaneously with the parameters of the underlying data model. Similar problems may exist in the selection model framework.⁶²

15.4 LONGITUDINAL ANALYSIS OPTIONS FOR “COMPLETE” DATA SETS

In general the methods described below are applicable to both repeated measures on an individual over time and measurements of different scales or scores on a given individual at the same point in time. Many studies, of course, utilize both of these

designs, asking patients to fill out questionnaires comprising several subscales at repeated intervals over the course of the study.

15.4.1 NORMAL LINEAR MIXED MODEL

The normal linear mixed model⁶³ is widely used in longitudinal analysis. For $i = 1, \dots, n$ individuals, the repeated QOL measurements are organized into outcome vectors y_i , which are assumed to be normally distributed, and a linear function of both fixed effects X_i and random effects Z_i , as follows:

$$y_i = X_i\alpha + Z_ib_i + e_i$$

where α is a vector of parameters linking the outcomes to the fixed effects, the random parameters b_i follow k -variate normal distributions with mean zero and variance-covariance matrix D , and the e_i follow multivariate normal distributions with mean zero and variance-covariance matrix S_i ; here D is a positive-definite covariance matrix of dimension k , and S_i is a positive-definite covariance matrix of dimension n_i whose parameters do not depend on i . The model can be formulated using either a classical likelihood approach or Bayesian methodology; in both cases the EM algorithm⁶⁴ can be used for estimation. When the variance parameters in S_i and D are unknown, they can be estimated using either maximum likelihood (ML) or restricted maximum likelihood (REML). REML estimates avoid the downward bias of the ML estimates, which occurs because they fail to account for the degrees of freedom lost in estimating the regression parameters. Software is available to fit this model in many standard packages, including SAS and Splus.^{65, 66}

The linear mixed model is very appealing, as it allows for individual effects on QOL in addition to fixed effects for treatment, time, and clinical variables. These effects can describe individual variation around the average outcome at different times, or they can take the form of growth curve parameters for individuals. Further, more traditional analysis of variance (ANOVA) and covariance (ANCOVA) models are simply special cases of the model stated here in which the $b_i = 0$.

15.4.2 GENERALIZED LINEAR MODELS

A second general class of models is the likelihood-based generalized linear model (GLM).⁶⁷ This framework is attractive because it accommodates a whole class of data rather than being restricted to continuous Gaussian measurements; it allows a unified treatment of measurements of different types with specification of an appropriate link function that determines the form of the mean and variance. For example, binary data can be evaluated using a logistic link function in order to evaluate the proportion of subjects experiencing a particular outcome. Estimation proceeds by solving the likelihood score equations, usually using iteratively reweighted least squares or Newton–Raphson algorithms. GLMs can be fit with GLIM,⁶⁸ with Splus using the GLM function,⁶⁹ or with SAS using the GLM or mixed procedures.⁷⁰ If missing data are MAR, unbiased estimates will be obtained. Generalized linear mixed models are a useful extension allowing for the inclusion of random effects in the GLM framework. SAS macros are available to fit these models.⁷¹

15.4.3 GENERALIZED ESTIMATING EQUATIONS

Like GLMs, generalized estimating equations (GEEs)⁷² provide a framework to treat disparate kinds of data in a unified way. In addition, they require specification of only the first two moments of the repeated measures rather than the likelihood. Estimates are obtained by solving an estimating equation of the following form:

$$U = \sum_{i=1}^n D_i' V_i^{-1} (Y_i - \mu_i) = 0$$

Here $\mu_i = E(Y_i | X_i, \beta)$ and $D_i = \partial \mu_i / \partial \beta$ are the usual mean and derivative functions, and V_i is a working correlation matrix. For Gaussian measurements, the estimating equations resulting from the GEE are equivalent to the usual score equations obtained from a multivariate normal maximum likelihood model; the same estimates will be obtained from either method. Software is available in the form of a SAS macro.⁷³

15.4.4 CHANGE-SCORE ANALYSIS

Analysis of individual or group changes in QOL scores over time is often of great importance in longitudinal studies. The simplest and most commonly used type of change-score analysis is to take the difference of QOL outcomes at two time points and apply a t-test or non-parametric test to compare treatment groups. Alternatively, one can assess the extent of change by applying a paired t-test to repeated measurements within a population. Change-score analysis has the advantage of inherently adjusting for the baseline score but must also be undertaken with caution as it is by nature sensitive to problems of regression to the mean.⁷⁴ In QOL studies in particular, a large amount of individual subject variation can overwhelm a statistically significant but small effect; changes in means scores by treatment group must be interpreted with great caution before being applied to individual patients.⁷⁵

15.4.5 TIME-TO-EVENT ANALYSIS

If attainment of a particular QOL score or milestone is the basis of the experiment, time-to-event or survival analysis methods can be applied. Once the event has been clearly defined, the analysis tools can be directly applied. These include Kaplan–Meier estimates of survival functions,⁷⁶ Cox proportional hazard regression models⁷⁷ to relate covariates to the probability of the event, and logrank and other tests for differences in the event history among comparison groups. The QOL database, however, supports few such milestones at this time.

15.5 METHODS FOR ANALYSIS WITH MISSING DATA

15.5.1 JOINT MODELING OF MEASUREMENT AND MISSINGNESS PROCESSES

One can model the joint distribution of the underlying complete data Y_i and the missingness indicators R_i . If conditioning arguments are used, two types of models can result; the selection model is concerned with $f(Y_i)f(R_i | Y_i)$, while the pattern mixture model is concerned with $f(R_i)f(Y_i | R_i)$. The two approaches are discussed and

compared in detail by Little.⁶¹ Pattern mixture models proceed by estimating the parameters of interest within strata defined by patterns of and/or reasons for missingness and then combining the estimates. Selection models proceed by modeling the complete data and then modeling the behavior of the missingness probabilities conditional on the outcome data.

Selection models for continuous and event data have been used by many authors; see the following for application to QOL.^{62, 78, 79, 80} While the computations can be burdensome, the approach will produce unbiased estimates even in the face of MNAR processes, provided that both parts of the model are correctly specified. Most selection models assume that the complete underlying responses are multivariate normal; any parametric model, such as the logistic, can be used for the missing data probabilities. The type of missingness mechanism is controlled by the covariates and/or responses that are included in the model for the missingness probabilities. For example, Troxel et al.⁷⁸ posit a multivariate normal model for coping scores in a breast cancer trial, assuming an autoregressive covariance structure, along with a logistic model for the missing data probabilities that depends on current and possibly unobserved values of QOL:

$$Y_{it}^* | Y_{i,t-1}^* \sim N \left\{ \mu_{it} + \rho_{t-1} \frac{\sigma_t}{\sigma_{t-1}} (Y_{i,t-1}^* - \mu_{i,t-1}), \sigma_t^2 (1 - \rho_{t-1}^2) \right\}$$

$$\pi_{it} = \text{logit}[P(R_{it} = 0)] = \beta_0 + \beta_1 Y_{it}^*$$

where Y_{it}^* represents the true but potentially unobserved QOL score for subject i at time t , $\sigma_t^2 = \text{var}(Y_{it}^*)$, $\rho_t = \text{corr}(Y_{it}^*, Y_{i,t-1}^*)$, R_{it} is an indicator for subject i at time t taking the value 1 if Y_{it}^* is observed and 0 otherwise, and π_{it} is the probability that Y_{it}^* is missing. Since the probabilities depend on the current, possibly unobserved measurement Y_{it}^* , the model can handle MNAR data; it is possible to allow dependence on previous values as well. The observed data likelihood is obtained by integrating the complete data likelihood over the missing values, as follows:

$$\int \cdots \int f_{i1}^* \pi_{i1}^{1-R_{i1}} (1 - \pi_{i1})^{R_{i1}} \prod_{t=2}^T f_{it-1}^* \pi_{it-1}^{1-R_{it-1}} (1 - \pi_{it-1})^{R_{it-1}} \prod_{t=1}^T (dY_{it}^*)^{1-R_{it}}$$

where f_{i1}^* is the normal density for Y_{i1}^* and f_{it-1}^* is the conditional normal density for $Y_{it}^* | Y_{i,t-1}^*$; π_{i1} is the probability that the first value for subject i is observed; and T is the last assessment time. Estimates are usually obtained through direct maximization of the likelihood surface; numerical integration is generally required. Once estimates are obtained, inference is straightforward using standard likelihood techniques. This method allows analysis of all the data, even when the missingness probabilities depend on potentially unobserved values of the response. The estimates are also likely to depend on modeling assumptions, most of which are untestable in the presence of MNAR missing data. Despite these drawbacks, these models can be very useful for investigation and testing of the missingness mechanism. In addition, the bias that results from assuming the wrong type of missingness mechanism may well be more severe than the bias that results from misspecification of a full maximum likelihood model. Software to fit the Diggle and Kenward⁶² model is available.⁸¹

Many authors have proposed pattern-mixture models.^{82–85} Pauler et al.⁸⁴ used a pattern-mixture model approach to model repeated measures conditional on death and drop-out times with a multinomial model for the death times themselves. The QOL outcomes follow a normal linear mixed model where

$$Y_i(t) | (b_{0i}, b_{1i}) \sim N(\beta_0 + \beta_1 t + \beta_2 z + \beta_3 zt + b_{0i} + b_{1i}t, \sigma^2)$$

$$(b_{0i}, b_{1i})' \sim N_2((0,0)', D)$$

and t and z are time and treatment effects, respectively; b_{0i} and b_{1i} are individual-specific intercepts and slopes; and σ^2 is the conditional within-subject variation.

For discrete data, methods allowing for nonignorable missing data have also been proposed.^{86–91} Often, log-linear models are used for the joint probability of outcome and response variables conditional on covariates. The models can be fit using the EM algorithm⁶⁴ treating the parameters of the missingness model as a nuisance or using estimating equations.

Several extensions to standard mixed models have been proposed in the context of longitudinal measurements in clinical trials. Zee⁹² has proposed growth curve models where the parameters relating to the polynomial in time are allowed to differ according to the various health states experienced by the patient, for example, on treatment, off treatment, and post-relapse. This method requires that the missing data be MAR, and may be fit with standard packages by simply creating an appropriate variable to indicate health state; in essence it is a type of pattern mixture model.

Schluchter⁹³ has proposed a joint mixed effects model for the longitudinal assessments and the time to dropout in which a vector of random effects and the log of the dropout time are jointly normally distributed. This model allows MNAR data in the sense that the time of dropout is allowed to depend on the rate of change in the underlying measurements. More recently, a number of authors have proposed sophisticated joint models for longitudinal and event-time data. Work in HIV/AIDS focused on repeated measurements of CD4 counts as predictors of recurrence or other kinds of failure and posited a measurement error model for the repeated measurements and a Cox model for the survival outcome.^{94–97} Because these models account for measurement error in the repeated measures, they are readily adaptable to longitudinal values with missing data; in general the data must be MAR for valid inference, but they can be used for MNAR mechanisms if the dropout times are determined by survival data. More recent developments include estimation via estimating equations,⁹⁸ relaxation of assumptions of normality,⁹⁹ and use of Bayesian methods to handle longitudinal measures with varying assessment schedules.¹⁰⁰

15.5.2 WEIGHTED GEEs

GEEs produce unbiased estimates for data that are MCAR. Extensions to the GEE exist for data that are MAR; weighted GEEs will produce unbiased estimates provided the weights are estimated consistently.^{101, 102} When the missingness probabilities depend only on observed covariates, such as the stage of disease, or responses, such as the baseline QOL score, a logistic or probit model can be used to estimate

missingness probabilities for every subject; the weights used in the analysis are then the inverses of these estimated probabilities. Robins et al.¹⁰² discuss these equations and their properties in detail; presented simply, the estimating equation takes the form

$$U = \sum_{i=1}^n D_i' V_i^{-1} \text{diag}\left(\frac{R_i}{\hat{\pi}_i}\right) (Y_i - \mu_i) = 0$$

where $\pi_i = P(R_{ij} = 1 \mid Y_i^0, W_p, \alpha)$, $\hat{\pi}_i$ is an estimate of π_i , and $\text{diag}(Q)$ indicates a matrix of zeroes with the vector Q on the diagonal. Although software exists to fit GEEs, additional programming is required to fit a weighted version.

15.5.3 SENSITIVITY ANALYSIS

Even while more sophisticated methods for handling missing data are developed, sensitivity analysis remains an integral part of any analysis involving incomplete data. Sensitivity analyses, in which the parameters governing the missing data mechanism are varied to determine their effect on estimates of the parameters of interest, are even more crucial in settings where nonignorable missing data may occur.

Sensitivity analyses can take several forms. Some authors have recommended global sensitivity analyses in the context of selection models for likelihood-based⁵⁶ or semiparametric inference.¹⁰³ Others have proposed analogous approaches in the context of pattern-mixture models.^{104,105} These analyses can be very useful but can sometimes still involve assumptions about the complete data distribution and require difficult computations. Another option is to study sensitivity locally in the neighborhood of the MAR model. Copas and Li¹⁰⁶ describe an approach for a normal linear model in which the correlation of error terms in the model of interest and the missing data model is treated as a sensitivity parameter. Others have suggested using ideas of local influence and likelihood displacement to assess sensitivity due to individual observations or collections of observations.¹⁰⁷ More recently, Troxel et al.¹⁰⁸ proposed a more general index of local sensitivity in the context of parametric selection models. This approach focuses on the behavior of the maximum likelihood estimate (MLE) as the nonignorability parameter moves away from its value under the MAR model (usually zero). It requires no specialized computations and focuses on the neighborhood of the MAR model because that is the preferred mode of analysis if the sensitivity is not extreme.

15.6 AVOIDING PITFALLS: SOME COMMONLY USED SOLUTIONS

15.6.1 SUBSTITUTION METHODS

In general, methods that rely on substitution of some value, determined in a variety of ways, are subject to bias and heavily subject to assumptions made in obtaining the substituted value. For these reasons, they should not be used to produce a primary analysis on which treatment or other decisions are based. One problem with substitution

methods, especially when the worse score method is used, is that they can seriously affect the psychometric properties of a measure. These properties, such as reliability and validity, rely on variations in the scores in order to hold. A second problem is that in substituting values and then conducting analyses based on those data, the variance of estimates will be underestimated because the missing values, had they been observed, would carry with them random variation, which the substituted values do not. Substitution methods can be useful, however, in conducting sensitivity analyses to determine the extent to which the analysis is swayed by differing data sets.

The worse score method is often used in sensitivity analyses, making the implicit assumption that subjects who did not submit an assessment are as badly off as they can possibly be with respect to QOL. This is usually an extreme assumption, so an analysis robust to worst score substitution has a strong defense.

The last value carried forward (LVCF) substitution method tries to use each patient's score to provide information about the imputed value. It assumes, however, that subjects who drop out do not have a changing QOL score, when in practice the subjects in rapid decline often tend to drop out prematurely. For this reason LVCF should be used with extreme care if at all.

Use of the average score either within a patient or within a group of patients, such as those on the same treatment arm, is more closely related to classic imputation methods. Again, it assumes that the imputed values are no different from the observed values, but it does not necessarily force each subject's score to remain constant.

15.6.2 MULTIPLE IMPUTATION

Imputation, or "filling-in," of data sets is a valid way of converting an incomplete data set to a complete data set. This method is attractive because once the imputation is conducted the methods for complete data described in Section 15.4 can be applied. Unlike ordinary substitution, single imputation consists of substituting a value for the missing observations and then adjusting the analysis to account for the fact that the substituted value was not obtained with the usual random variation.

Multiple imputation¹⁰⁹ is similar in spirit to single imputation but with added safeguards against underestimation of variance due to substitution. Several data sets are imputed, and the analysis in question is conducted on each of them resulting in a set of estimates obtained from each imputed data set. These several results are then combined to obtain final estimates based on the multiple sets.

Multiple imputation can be conducted in the presence of all kinds of missingness mechanisms. The usual drawback with respect to nonignorable missingness applies, however. A model is required to obtain the imputed values, and in the presence of nonignorable missingness the resultant estimates are sensitive to the chosen model; even worse, the assumptions governing that model are generally untestable.

15.6.3 ADJUSTED SURVIVAL ANALYSES

Some authors have proposed analyses in which survival is treated as the primary outcome with adjustment for the QOL experience of the patients. This is an extremely appealing idea, for it clarifies the inherent trade-off between length and quality of life that applies to most patients. It can be difficult to implement satisfactorily in practice,

however, because of the difficulty of obtaining the appropriate weights for survival in different periods. The two methods described below have gained some popularity.

15.6.3.1 Quality Adjusted Life Years

This method consists of estimating a fairly simple weighted average in which designated periods of life are weighted according to some utility describing QOL.^{110–114} Because utilities are obtained using lengthy interviews or questionnaires focusing on time trade-offs or standard gambles, investigators commonly substitute utilities obtained from some standard population rather than any information obtained directly from the patient. This renders the analysis largely uninterpretable in our view.

15.6.3.2 Q-TWiST

Q-TWiST,^{115–119} or quality-adjusted time without symptoms and toxicity, is a more detailed method of adjustment, though still one that relies on utilities. The patient's course through time is divided into intervals in which the patient experiences toxicity due to treatment, toxicity due to disease brought on by relapse, and no toxicity. These intervals may be somewhat arbitrary, determined not by the patient's actual experience with toxicity but by predefined averages observed among patients with a given disease receiving a given treatment. To compound this arbitrariness, utilities for each period are chosen by the analyst, but the impact of different utilities can be examined in sensitivity analyses. This results in an analysis that reflects only a small amount of patient-derived data and a large number of parameters chosen by the investigator. Data from patient rating scales and Q-TWiST analyses can differ.¹²⁰

REFERENCES

1. Moinpour CM, Feigl P, Metch B, Hayden KH, Meyskens FL Jr, Crowley J. Quality of life end points in cancer clinical trials: review and recommendations. *J Natl Cancer Inst* 1989;81:485–495.
2. Nayfield S, Ganz PA, Moinpour CM, Cella D, Hailey B. Report from a National Cancer Institute (USA) workshop on quality of life assessment in cancer clinical trials. *Quality of Life* 1992;1:203–210.
3. National Cancer Institute (US). Quality of life in clinical trials. Proceedings of a workshop held at the National Institute of Health; March 1–2, 1995. Bethesda, MD: National Institute of Health, 1996.
4. Sugarbaker PH, Barofsky I, Rosenberg SA, Gianola FJ. Quality of life assessment of patients in extremity sarcoma clinical trials. *Surgery* 1982;17–23.
5. Nunnally, J. *Psychometric Theory*. New York: McGraw-Hill, 1978.
6. Kirshner B, Guyatt GH. A methodological framework for assessing health indices. *J Chron Dis* 1985;1:27–36.
7. Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. *J Clin Epidemiol* 1989; 42:403–408.

8. Cella DF, Tulsky DS, Gray G, Sarafian B, Linn E, Bonomi AE, et al. The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. *J Clin Oncol* 1993; 11:570–579.
9. Cella DF, Bonomi AE, Lloyd ST, Tulsky DS, Kaplan E, Bonomi P. Reliability and validity of the Functional Assessment of Cancer Therapy — Lung (FACT-L) quality of life instrument. *Lung Cancer* 1995; 12:199–220.
10. Esper P, Mo F, Chodak G, Sinner M, Cella D, Pienta KJ. Measuring quality of life in men with prostate cancer using the Functional Assessment of Cancer Therapy — Prostate instrument. *Urology* 1997; 50:920–928.
11. Cella DF. Manual of the Functional Assessment of Chronic Illness Therapy (FACIT Scales) — Version 4. Outcomes Research and Education (CORE). 4. Evanston Northwestern Healthcare and Northwestern University, 1997.
12. Webster K, Odom L, Peterman A, Lent L, Cella D. The Functional Assessment of Chronic Illness Therapy (FACIT) measurement system: validation of version 4 of the core questionnaire. *Qual Life Res* 1999;8:604.
13. Aaronson NK, Bullinger M, Ahmedzai S. A modular approach to quality of life assessment in cancer clinical trials. *Recent Results Cancer Res* 1988;111:231–249.
14. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, Filiberti A, Flechtner H, Fleishman SB, de Haes JCJM, Kaasa S, Klee M, Osoba D, Razavi D, Rofe PB, Schraub S, Sneeuw K, Sullivan M, Takeda F for the European Organization for Research and Treatment of Cancer Study Group on Quality of Life. 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.
15. Fayers P, Aaronson N, Bjordal K, Groenvold M, Curran D, Bottomley A on behalf of EORTC Quality of Life Study Group. *EORTC QLQ-C30 Scoring Manual*. 3rd ed. Brussels: EORTC, 2001.
16. Fayers P, Weeden S, Curran D on behalf of EORTC Quality of Life Study Group. *EORTC QLQ-C30 Reference Values*. Brussels, Belgium: EORTC Quality of Life Study Group, 1998.
17. Fayers P, Bottomley A on behalf of the EORTC Quality of Life group and of the Quality of Life Unit. Quality of life research within the EORTC: the EORTC QLQ-C30. *Eur J Cancer* 2002;38:S125–S133.
18. www.facit.org/qview/qlist.aspx (accessed August 2004).
19. www.eortc.be/home/qol (accessed August 2004).
20. Schag CAC, Ganz PA, Heinrich Richard L. Cancer rehabilitation evaluation system — Short Form (CARES-SF): a cancer specific rehabilitation and quality of life instrument. *Cancer* 1991;68:1406–1413.
21. Heinrich RL, Schag CAC, Ganz PA. Living with cancer: the cancer inventory of problem situations. *J Clin Psychol* 1984;40:972–980.
22. Schag CAC, Heinrich RL, Ganz PA. The cancer inventory of problem situations: an instrument for assessing cancer patients' rehabilitation needs. *J Psychosocial Oncol* 1983;1:11–24.
23. Schag CAC, Heinrich RL. Development of a comprehensive quality of life measurement tool: CARES. *Oncology* 1990;4:135–138.
24. Stewart AL, Ware JE Jr. *Measuring Functioning and Well Being: The Medical Outcomes Approach*. Durham and London: Duke University Press, 1992.
25. Ware JE Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36): I. conceptual framework and item selection. *Med Care* 1992;30:473–483.

26. McHorney C, Ware JE Jr., Raczek A. The MOS 36-item short-form health survey (SF-36): II. psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med Care* 1993;31:247–263.
27. McHorney CA, Ware JE Jr., Lu RJF, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): III. tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Med Care* 1994;32:40–46.
28. Ware JE Jr., Snow KK, Kosinski MA, Gandek B. *SF-36 Health Survey: Manual and Interpretation Guide*. Boston: Nimrod Press, 1993.
29. Ware JE Jr., Kosinski M, Keller SD. *SF-36 Physical and Mental Health Summary Scales: A User's Manual*. Boston: The Health Institute, New England Medical Center, 1994.
30. Ware JE Jr. The SF-36 Health Survey. In Spilker B, ed. *Quality of Life and Pharmacoeconomics in Clinical Trials*. Philadelphia: Lippincott-Raven, 1996.
31. Ware JE, Kosinski M, Dewey JE. How to score version two of the SF-36 Health Survey. Lincoln, RI: QualityMetric, Incorporated, 2000.
32. Ware JE, Kosinski M, Turner-Bowker DM, Gandek B. How to score version 2 of the SF-12® Health Survey (with a supplement documenting version 1). Lincoln, RI: QualityMetric Incorporated, 2002.
33. Turner-Bowker DM, Bartley PJ, Ware JE Jr. *SF-36® Health Survey & "SF" Bibliography*. 3rd ed. (1998–2000). Lincoln, RI: QualityMetric Incorporated, 2002.
34. Ware JE Jr., Kosinski M, Dewey JE, Gandek B. *How to Score and Interpret Single-Item Health Status Measures: A Manual for Users of the SF-8™ Health Survey*. Lincoln, RI: QualityMetric Incorporated, 2001.
35. Turner-Bowker DM, Bayliss MS, Ware JE Jr., Kosinski M. Usefulness of the SF-8™ Health Survey for comparing the impact of migraine and other conditions. *Qual Life Res* 2003;12:1003–1012.
36. www.sf-36.org (accessed August 2004).
37. Moinpour CM, Savage MJ, Troxel AB, Lovato LC, Eisenberger M, Veith RW, Higgins B, Skeel R, Yee M, Blumenstein BA, Crawford ED, Meyskens FL Jr. Quality of life in advanced prostate cancer: results of a randomized therapeutic trial. *J Natl Cancer Inst* 1998;90:1537–1544.
38. Tang ST, McCorkle R. Appropriate time frames for data collection in quality of life research among cancer patients at the end of life. *Qual Life Res* 2002; 11:145–155.
39. Hakamies-Blomqvist L, Luoma M-L, Sjöström J, Pluzanska A, Sjödin M, Mouridsen H, Østenstad B, Mjaaland I, Ottosson S, Gergth J, Malmström P-O, Blomqvist C. Timing of quality of life (QoL) assessments as a source of error in oncological trials. *J Advanced Nurs* 2001;35:709–716.
40. Sprangers MAG, Moinpour CM, Moynihan TJ, Patrick DL, Revicki DA, and the Clinical Significance Consensus Meeting Group. Assessing meaningful change in quality of life over time: a user's guide for clinicians. *Mayo Clin Proc* 2002; 77:561–571.
41. Sloan JA, Cella D, Frost M, Guyatt GH, Springers M, Symonds T, and the Clinical Significance Consensus Meeting Group. Assessing clinical significance in measuring oncology patient quality of life: introduction to the symposium, content overview, and definition of terms. *Mayo Clin Proc* 2002;77:367–370.
42. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR, and the Clinical Significance Consensus Meeting Group. Methods to explain the clinical significance of health status measures. *Mayo Clin Proc* 2002;77:371–383.

43. Cella D, Bullinger M, Scott C, Barofsky I, and the Clinical Significance Consensus Meeting Group. Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life. *Mayo Clin Proc* 2002; 77:384–392.
44. Sloan JA, Aaronson N, Cappelleri JC, Fairclough DL, Varricchio C, and the Clinical Significance Consensus Meeting Group. Assessing the clinical significance of single items relative to summated scores. *Mayo Clin Proc* 2002;77:479–487.
45. Frost MH, Bonomi AE, Ferrans CE, Wong GY, Hays RD, and the Clinical Significance Consensus Meeting Group. Patient, clinician, and population perspectives on determining the clinical significance of quality-of-life scores. *Mayo Clin Proc* 2002;77:488–494.
46. Symonds T, Berzon R, Marquis P, Rummans TA, and the Clinical Significance Consensus Meeting Group. The clinical significance of quality-of-life results: practical considerations for specific audiences. *Mayo Clin Proc* 2002;77:572–583.
47. Cella D, Eton DT, Fairclough DL, Bonomi P, Heyes AE, Silberman C, Wolf MK, Johnson DH. What is a clinically meaningful change on the Functional Assessment of Cancer Therapy — Lung (FACT-L) Questionnaire? Results from Eastern Cooperative Oncology Group (ECOG) Study 5592. *J Clin Epidemiol* 2002; 55:285–295.
48. King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. *Qual Life Res* 1996;5:555–567.
49. Osoba D, Rodriques G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. *J Clin Oncol* 1998;16:139–144.
50. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
51. Sloan JA, Loprinzi CL, Kross SA, Miser AW, O’Fallon JR, Mahoney MR, Heid IM, Bretscher ME, Vaught NL. Randomized comparison of four tools measuring overall quality of life in patients with advanced cancer. *J Clin Oncol* 1998;16:3662–3673.
52. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? *J Amer Med Assoc* 2003;289:2554–2559.
53. Johnson JR, Williams G, Pazdur R. End points and United States Food and Drug Administration approval of oncology drugs. *J Clin Oncol* 2003;21:1404–1411.
54. Rubin DB. Inference and missing data. *Biometrika* 1976;63:581–592.
55. Machin D, Weeden S. Suggestions for the presentation of quality of life data from clinical trials. *Stat Med* 1998;17:711–724.
56. Troxel AB. A comparative analysis of quality of life data from a Southwest Oncology Group randomized trial of advanced colorectal cancer. *Stat Med* 1998;17:767–779.
57. Curran D, Bacchi M, Schmitz SFH, Molenberghs G, Sylvester RJ. Identifying the types of missingness in quality of life data from clinical trials. *Stat Med* 1998; 17:697–710.
58. Diggle PJ. Testing for random dropouts in repeated measurement data. *Biometrics* 1989;45:1255–1258.
59. Ridout M. Reader reaction: testing for random dropouts in repeated measurement data. *Biometrics* 1991;47:1617–1621.
60. Molenberghs G, Goetghebeur EJT, Lipsitz SR, Kenward MG. Nonrandom missingness in categorical data: strengths and limitations. *Amer Statist* 1999;53:110–118.
61. Little RJA. Modeling the drop-out mechanism in repeated-measures studies. *J Am Stat Assoc* 1995;90:1112–1121.
62. Diggle P, Kenward, M. Informative drop-out in longitudinal analysis (with discussion). *Appl Stat* 1994;43:49–93.

63. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; 38:963–974.
64. Dempster AP, Laird NM, Rubin DB. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J Roy Statist Soc B* 1977; 39:1–38.
65. The SAS Institute, Cary NC; www.sas.com/index.html (accessed August 2004).
66. Insightful Corporation, Seattle WA; www.insightful.com/products/splus/default.asp (accessed August 2004).
67. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd ed. London: Chapman and Hall, 1989.
68. Baker RJ, Nelder JA. The GLIM System, Release 3, Generalized Linear Interactive Modeling. Oxford: Numerical Algorithms Group, 1978.
69. Hastie TJ, Pregibon D. Generalized linear models. In Chambers JM, Hastie TJ, eds. *Statistical Models in S*. London: Chapman and Hall, 1993.
70. Wolfinger R, Chang M. *Comparing the SAS GLM and MIXED Procedures for Repeated Measures*. Cary, NC: The SAS Institute.
71. ftp.sas.com/techsup/download/stat/glmm800.html (accessed August 2004).
72. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986;73:13–22.
73. Groemping U. GEE: a SAS macro for longitudinal data analysis. Technical report, Fachbereich Statistik, Universitaet Dortmund, Germany, 1994.
74. Fleiss JL. *The Design and Analysis of Clinical Experiments*. New York: John Wiley and Sons, 1986.
75. Donaldson GW, Moinpour CM. Individual differences in quality-of-life treatment response. *Med Care* 2002;40:III-39–III-53.
76. Kaplan EL, Meier P. Nonparametric estimator from incomplete observations. *J Amer Statist Assoc* 1958;53:457–481.
77. Cox, DR. Regression models and life tables (with discussion). *J Roy Statist Soc B* 1972;21:411–421.
78. Troxel AB, Harrington DP, Lipsitz SR. Analysis of longitudinal data with non-ignorable non-monotone missing values. *Appl Stat* 1998;47:425–438.
79. Beacon HJ, Thompson SC. Multi-level models for repeated measurement data: application to quality of life in clinical trials. *Stat Med* 1996;15:2717–2732.
80. Michiels B, Molenberghs G, Bijlens L, Vangeneugden T, Thijs H. Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out. *Stat Med* 2002;21:1023–1041.
81. Smith DM. *The Oswald Manual*. Technical report, Statistics Group, University of Lancaster, Lancaster, England, 1996.
82. Little RJA. Pattern-mixture models for multivariate incomplete data. *J Amer Statist Assoc* 1993;88:125–134.
83. Little RJA. A class of pattern-mixture models for normal incomplete data. *Biometrika* 1994;81:471–483.
84. Pauler DK, McCoy S, Moinpour C. Pattern mixture models for longitudinal quality of life studies in advanced stage disease. *Stat Med* 2003;22:795–809.
85. Hogan JW, Laird NM. Mixture models for the joint distribution of repeated measures and event times. *Stat Med* 1997;16:239–257.
86. Fay RE. Causal models for patterns of nonresponse. *J Amer Statist Assoc* 1986; 81:354–365.
87. Baker SG, Laird NM. Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *J Amer Statist Assoc* 1988;83:62–69.

88. Conaway MR. The analysis of repeated categorical measurements subject to nonignorable nonresponse. *J Amer Statist Assoc* 1992;87:817–824.
89. Fitzmaurice GM, Molenberghs G, Lipsitz SR. Regression models for longitudinal binary responses with informative drop-outs. *J Roy Statist Soc B* 1995;57:691–704.
90. Lipsitz SR, Ibrahim JG, Fitzmaurice GM. Likelihood methods for incomplete longitudinal binary responses with incomplete categorical covariates. *Biometrics* 1999; 55:214–223.
91. Horton NJ, Fitzmaurice GM. Maximum likelihood estimation of bivariate logistic models for incomplete responses with indicators of ignorable and non-ignorable missingness. *App Stat* 2002;51:281–295.
92. Zee BC. Growth curve model analysis for quality of life data. *Stat Med* 1998; 17:757–766.
93. Schluchter MD. Methods for the analysis of informatively censored longitudinal data. *Stat Med* 1992;11:1861–1870.
94. Tsiatis AA, DeGruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error: applications to survival and CD4 counts in patients with AIDS. *J Amer Statist Assoc* 1995;90:27–37.
95. Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. *Biometrics* 1997;53:330–339.
96. Wang Y, Taylor JMG. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *J Amer Statist Assoc* 2001; 96:895–905.
97. Faucett CL, Schenker N, Elashoff RM. Analysis of censored survival data with intermittently observed time-dependent binary covariates. *J Amer Statist Assoc* 1998; 93:427–437.
98. Lipsitz SR, Ibrahim JG. Estimating equations with incomplete categorical covariates in the Cox model. *Biometrics* 1998;54:1002–1013.
99. Tsiatis AA, Davidian M. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* 2001; 88:447–458.
100. Pauler DK, Finkelstein DA. Predicting time to prostate cancer recurrence based on joint models for non-linear longitudinal biomarkers and event time outcomes. *Stat Med* 2002;21:3897–3911.
101. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *J Amer Statist Assoc* 1995;90:122–129.
102. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Amer Statist Assoc* 1995; 90:106–121.
103. Scharfstein DO, Rotnitzky A, Robin JM. Adjusting for nonignorable drop-outs using semiparametric nonresponse models. *J Amer Statist Assoc* 1999;94:1096–1120.
104. Little RJA, Wang Y. Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics* 1996;52:98–111.
105. Daniels MJ, Hogan JW. Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout. *Biometrics* 2000;56:1241–1248.
106. Copas JB, Li HG. Inference for non-random samples (with discussion). *J Roy Statist Soc B* 1997;59:55–95.
107. Verbeke G, Molenberghs G, Thijs H, Lesaffre E, Kenward MG. Sensitivity analysis for nonrandom dropout: a local influence approach. *Biometrics* 2001;57:7–14.
108. Troxel AB, Ma G, Heitjan DF. An index of local sensitivity to nonignorability. *Stat Sin* 2004; 14:1221–37.

109. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons, 1987.
110. Torrance GW. Designing and conducting cost-utility analyses. In B Spilker, ed. *Quality of Life and Pharmacoeconomics in Clinical Trials*. 2nd ed. Philadelphia: Lippincott-Raven Publishers, 1996, pp. 1105–1111.
111. La Puma J, Lawlor EF. Quality-adjusted life-years: ethical implications for physicians and policymakers. *J Amer Med Assoc* 1990;263:2917–2921.
112. Mehrez A, Gafni A. Quality-adjusted life years, utility theory, and healthy-years equivalents. *Med Decis Making* 1989;9:142–149.
113. Johannesson M. QALYs, HYE, and individual preferences: a graphical illustration. *Soc Sci Med* 1994;39:1623–1632.
114. Ramsey SD, Etzioni R, Troxel A, Urban N. Optimizing sampling strategies for estimating quality-adjusted life years. *Med Decis Making* 1997; 17:431–438.
115. Glasziou PP, Simes RJ, Gelber RD. Quality adjusted survival analysis. *Stat Med* 1990;9:1259–1276.
116. Gelber RD, Gelman RS. A quality-of-life-oriented endpoint for comparing therapies. *Biometrics* 1989;45:781–795.
117. Gelber RD, Goldhirsch A, Cole BF for the International Breast Cancer Study Group. Parametric extrapolation of survival estimates with applications to quality of life evaluation of treatments. *Control Clin Trials* 1993;14:485–499.
118. Goldhirsch A, Gelber RD, Simes J, Glasziou P, Coates A. Costs and benefits of adjuvant therapy in breast cancer: a quality-adjusted survival analysis. *J Clin Oncol* 1989; 7:36–44.
119. Gelber RD, Cole BF, Goldhirsch A. Comparing treatments using quality-adjusted survival: the Q-TWiST method. *Amer Statist* 1995;49:161–169.
120. Fairclough DL, Fetting JH, Cella D, Wonson W, Moinpour CM. Quality of life and quality adjusted survival for breast cancer patients receiving adjuvant therapy. *Qual Life Res* 2000;8:723–731.

16 Economic Analyses Alongside Cancer Clinical Trials

Scott D. Ramsey

CONTENTS

16.1	Introduction and Rationale.....	271
16.2	Design Issues for Economic Analyses Alongside Clinical Trials	273
16.2.1	Selecting Trials for Parallel Economic Evaluation	273
16.2.2	Statement of Study Objectives and Hypotheses	273
16.2.3	Sample Size and Power Estimation for Joint Clinical/ Economic Trials	274
16.3	Database Issues	275
16.3.1	Selecting Outcomes Endpoints	275
16.3.2	Extrapolating Costs and Outcomes Beyond the Trial Period	276
16.3.3	Measuring Costs.....	276
16.3.3.1	Tracking Resources Consumed in Care	277
16.3.3.2	Protocol-Induced Costs.....	277
16.3.3.3	Assigning Values to Resources Consumed.....	278
16.4	Analysis of Findings	278
16.4.1	Analysis of Costs	278
16.4.2	Per Protocol vs. Intent to Treat Analyses.....	279
16.4.3	Missing Data	279
16.4.4	Adjustment to Constant Dollars and Discounting of Future Costs and Outcomes	280
16.4.5	Summary Measures	281
16.4.6	Cost-Effectiveness Acceptability Curves	282
16.4.7	Net Health Benefit Approach.....	283
	References	284

16.1 INTRODUCTION AND RATIONALE

The ever-increasing cost of cancer care combined with the high prevalence and economic burden of cancer in economically developed societies has fostered a strong

demand for economic information regarding new technologies aimed at the prevention, detection, and treatment of cancer. As efforts to control health care spending become increasingly intense, decision makers are forced to confront the reality that adopting new, cost-increasing technologies necessitates spending less in other areas of health care. In this context, some have argued that it is reasonable to consider the outcomes and costs for cancer treatments relative to outcomes and costs for other medical interventions. The most common and accepted approach to compare the relative value of different interventions in creating better health and/or longer life is cost-effectiveness analysis (CEA).¹

Conducting cost-effectiveness analyses alongside clinical trials has two important advantages. First, it is an efficient and timely way to obtain data on clinical, economic, and humanistic outcomes simultaneously. Timely economic data will be particularly useful to those who are responsible for health care budgets and is often critical to formulary considerations. Second, performing a CEA alongside a randomized, controlled clinical trial has high internal validity and low potential for bias. Because economic considerations are unavoidable in clinical decision making, the highest quality economic evidence should be used.

Many countries now require economic information as part of applications for listing new products on national formularies. For example, the United Kingdom's National Institute for Clinical Excellence has published a guidance document for manufacturers and sponsors that includes information on submitting economic information as part of the application materials.² In the United States, the Academy of Managed Care Pharmacy (AMCP) has published its Format for Formulary Submissions, a guidance document for health insurance plans on requesting clinical and economic information on new medicines from pharmaceutical manufacturers.³ The AMCP Format has been widely adopted by health plans, and many pharmaceutical companies now routinely prepare AMCP style dossiers for new products prior to marketing.

To address the regulatory requirements of many countries and to provide timely and highly robust data on economic endpoints, economic analyses are now commonly performed alongside clinical trials of new treatments. The rising popularity of such piggyback economic studies poses an opportunity and a challenge for the field. Moving economic analyses in sync with clinical trials increases their value to decision makers and consequently their influence in clinical practice policy. The challenge is for researchers to raise the standards of these analyses so that they are considered of equal quality to the trials themselves. Although standard methods for conducting and reporting economic evaluations of health care technologies are available,^{4,5} developing a consistent approach for conducting them as part of clinical trials has only recently been recognized as a priority for the field.

Despite general agreement on principles for cost-effectiveness studies, there are concerns about the external validity of performing cost-effectiveness analyses alongside clinical trials. First, the clinical care that occurs in the trial is not representative of care that occurs in typical medical practice.^{6,7} This problem has several manifestations: (1) the control group often differs from standard practice, for example, placebo; (2) protocol-induced procedures such as eligibility screening tests artificially raise the cost of care; (3) strict screening and selection criteria mean that trial subjects are much more homogeneous and likely to comply with therapy than

patients typically seen in practice;⁴ investigators conducting the trial have particular expertise in the disease, and thus the results are likely to be better than will be achieved in typical practice settings.

Clinical trials and cost-effectiveness analyses are designed for different purposes. In short, clinical trials are designed to test whether a new intervention is efficacious (Does it work?), while cost-effectiveness analyses are designed to evaluate how the intervention should be prioritized in health care budgets (Is it worth the expenditure?). The distinction creates several potential conflicting issues: (1) clinical trials consider very narrowly defined patient populations that are cared for in a highly controlled and monitored environment, while CEAs consider the care of patients in clinical practice settings; (2) outcome measures for trials are typically surrogate clinical endpoints, for example, serum cholesterol, while appropriate outcome measures for CEAs include survival and health-related quality of life; (3) sample size needs for clinical endpoints may be insufficient for joint clinical/economic trials; (4) the time horizon for clinical trials is usually shorter than the relevant time horizon for CEAs, which considers the course of disease.

Including CEA-related endpoints increases the cost of data collection in a clinical trial. It can be difficult and expensive to design forms and collect data on health care utilization, both trial and nontrial related, and costs. In addition, tracking CEA-related endpoints, such as survival, may entail enlarging the sample size or extending the period of observation beyond that which is often necessary to establish clinical effectiveness. Such additional costs can be difficult to justify when the outcome of the trial is uncertain.

16.2 DESIGN ISSUES FOR ECONOMIC ANALYSES ALONGSIDE CLINICAL TRIALS

16.2.1 SELECTING TRIALS FOR PARALLEL ECONOMIC EVALUATION

Dozens of cancer-related controlled clinical trials are started each year. Not all of them warrant the cost and expense necessary to support a parallel economic evaluation. Given research budget constraints, it is important to be selective and provide adequate resources for economic analyses.

16.2.2 STATEMENT OF STUDY OBJECTIVES AND HYPOTHESES

Most clinical trials are designed with the goal of assessing the efficacy of an experimental intervention compared to a standard treatment or placebo. Thus, a null hypothesis is set forth: $H_0: \mu_a = \mu_b$ vs. an alternative $H_1: \mu_a \neq \mu_b$, where μ is a clinical outcome of interest, such as blood pressure, and a and b are the competing therapies. In contrast, cost-effectiveness studies are designed to estimate the incremental cost-effectiveness of two alternative interventions for a given medical condition, calculated as follows:

$$\text{Incremental Cost-Effectiveness}_a = R = \frac{(Cost_a - Cost_b)}{(Effectiveness_a - Effectiveness_b)} = \frac{\Delta C}{\Delta E} \quad (16.1)$$

In the context of cost-effectiveness studies alongside randomized trials, treatment a is the experimental therapy and treatment b is the standard therapy for the condition. The most appropriate comparison intervention for a CEA ($Effectiveness_b$) is the most effective therapy or therapies for the condition, but unfortunately the comparison for a clinical trial is often placebo in many cases. $Cost_a$ and $Cost_b$ are the average costs per patient for persons in the intervention and control groups, respectively, and include the cost of therapy and other related medical care such as side effects related to chemotherapy for the condition. $Effectiveness$ can be measured in ad hoc units, such as relapse-free survival, or in common metrics that facilitate comparison across interventions. The most common metrics of effectiveness are life years or quality-adjusted life years.

In a seminal article, O'Brien and colleagues⁹ suggested that the appropriate null and alternative hypotheses for a clinical trial evaluating cost-effectiveness should be $H_0: R = R_{max}$ vs. $H_1: R < R_{max}$, where R is the incremental cost-effectiveness of the experimental intervention, given by Equation (16.1), and R_{max} is a prespecified threshold. A persistent issue that haunts cost-effectiveness researchers is that there is no widely agreed upon R_{max} . Indeed, it is likely that R_{max} varies substantially among organizations and governments that fund medical care. As will be discussed below, this issue is particularly important because the value selected for R_{max} will greatly affect the sample size requirements for the study.

16.2.3 SAMPLE SIZE AND POWER ESTIMATION FOR JOINT CLINICAL/ECONOMIC TRIALS

With regard to sample size and the appropriate threshold (R_{max}) for hypothesis testing, budgetary factors will usually dictate the number of subjects that can be enrolled in the clinical trial. It may be reasonable to take sample size as given and solve for R_{max} using the usually accepted levels of power ($1-\beta$) and significance (α). Decision makers can then decide if the study will achieve adequate power to evaluate a cost-effectiveness threshold that is meaningful to their organization. There are certain technical challenges to calculating the power curves for cost-effectiveness studies because more parameters must be solicited for clinical/economic trials than for typical clinical trials.

Classical RCTs without economic end points often have a single primary end point. For example, an RCT for cancer treatment may monitor 5-year survival past cancer treatment with an outcome $Y = 1$ if alive at 5 years and 0 if dead at 5 years. The effect size estimate the difference or ratio of the survival rates in treatment, $P_1 = E(Y_i | \text{Treatment})$, and control, $P_0 = E(Y_i | \text{Control})$. With economic outcomes, the difference is preferred as an effect size estimator. Power calculations for a traditional RCT are a matter of determining an adequate sample size to assure that the difference in P_1 and P_0 can be observed.

Economic studies measure the effect of the treatment on the total dollars expended. If D refers to the total treatment dollar expenditure, then we could summarize the effect of treatment choice on costs alone by the difference in mean treatment costs $\mu_1 - \mu_0$ where $\mu_1 = E(D | \text{Treatment})$ and $\mu_0 = E(D | \text{Control})$. Evaluating the effect of treatment choice on dollars alone would require a simple power analysis to detect the difference. Solicitation for detecting differences in costs would typically also include stating $\sigma_1 = \text{StdDev}(D | \text{Treated})$ and $\sigma_1 = \text{StdDev}(D | \text{Control})$.

When economic and efficacy endpoints are combined, we summarize their joint behavior in the incremental cost-effectiveness ratio $R = (\mu_1 - \mu_0) / (p_1 - p_0)$. In addition to treatment effects on outcomes and costs, we must also characterize the *correlation* of costs and effects, $\rho = \text{cor}(D_i, Y_i)$ in order to calculate the sample size for the study. Although the correlation between costs and effects is not an explicit part of R or its estimator, $\hat{R} = (\hat{\mu}_1 - \hat{\mu}_0) / (\hat{p}_1 - \hat{p}_0)$, it does have great influence on the sampling behavior and hence power calculations. Positive correlations will tend to minimize variability compared to negative correlations.¹⁰ Unfortunately, soliciting the correlation between costs and outcomes is nonintuitive. Because the sampling behavior of $\hat{R}$ is most dependent on the correlation between costs and outcomes, power calculations for CEA should include sensitivity analyses when evaluating power. Relatively straightforward methods to determine power are available.¹¹

16.3 DATABASE ISSUES

16.3.1 SELECTING OUTCOMES ENDPOINTS

Possible health outcome measures for cost-effectiveness analyses include measures of averted morbidity, surrogate clinical markers, life extension, health-related quality of life, disability-adjusted survival, and quality-adjusted survival. Researchers have used treatment-specific measures such as symptom-free time, cases of disease prevented, and change in likelihood of progression of disease in the denominator of the cost-effectiveness ratio. The problem with these measures is that they are treatment- or disease-specific. Although treatment-specific effects make sense from a clinical perspective, from a resource allocation perspective, the health policy analyst would prefer to have a common and standard measure of health outcome that extends across different diseases and treatments to facilitate comparability of cost-effectiveness studies.

Although there are no international standards on which outcome measure is most appropriate for cost-effectiveness evaluations, the U.S. Panel on Cost-Effectiveness in Medicine suggests using quality adjusted life years (QALYs) because it links to welfare theory and fits within a resource allocation and social utility framework.^{12,13} QALYs are defined as the duration of an individual's life as modified by the changes in health and well-being experienced over the lifetime.¹⁴ Well-being is measured as a health state utility value, ranging from 1 (ideal health) to 0 (death); [Figure 16.1](#) depicts this concept as the area under the curve.

There are several methods for obtaining health state utilities, including interviewer administered techniques and multiattribute survey methods.¹⁵ To obtain utility values, it is necessary to survey patients regularly over the course of the trial. In addition, if serious sequelae that influence quality of life occur sporadically, it may be necessary to sample patients experiencing those outcomes and survey them to obtain utilities for those states. This can become a complex task for clinical trialists wishing to keep the study procedures at a minimum. Finally, since trials are of finite duration while treatment effects related to cancer treatment can continue for a lifetime, it is important to consider how one will estimate these effects after the trial. Most cost-effectiveness analysts rely on modeling techniques or other indirect methods to estimate life expectancy and health state utilities after the trial has ended.¹⁶

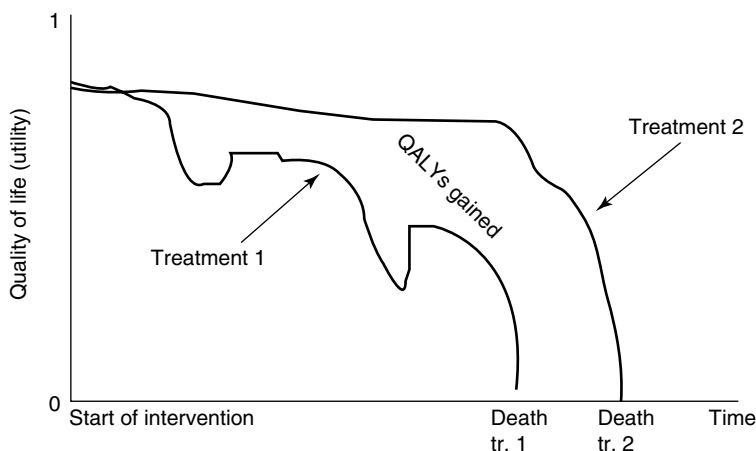


FIGURE 16.1 Quality-adjusted life years (QALYs) gained using a new therapy. QALYs are computed at the utility weights multiplied by survival for each period summed across all periods of interest. See Ramsey SD, Etzioni R, Troxel A, Urban N. Optimizing sampling strategies for estimating quality-adjusted life years. *Med Decis Making*. 1997;17:431–438. (Reprinted with permission from Ramsey SD, McIntosh M, Sullivan SD, *Annu Rev Public Health*, 2001; 22:129–41.)

16.3.2 EXTRAPOLATING COSTS AND OUTCOMES BEYOND THE TRIAL PERIOD

Often an intervention will affect health outcomes and costs long beyond the observation period of the clinical study. These costs and outcomes should be included in the cost-effectiveness analysis. If effects beyond the observation period are anticipated at the time of the trial, steps can be taken to address this issue. It may be useful to include in the study consent forms a notice that subjects will be contacted weeks, months, or years after the trial to determine how the intervention is affecting them. Arrangements can be made to collect billing records after the trial is completed, for example, Medicare claims. In most cases some degree of mathematical modeling will be necessary to project costs and outcomes over a plausible range of scenarios. Models can be reviewed prior to the trial to determine the most relevant application.^{17,18} Because many readers are skeptical of models and prefer to see results of the “real” data, presenting both the trial-period and lifetime cost-effectiveness ratios allows readers more flexibility in judging the outcomes of the trial.

16.3.3 MEASURING COSTS

Costs are a function of resources that are consumed during the course of care, multiplied by values or prices of each resource. Costs stemming from a therapy include the costs of:

1. Direct medical care, including the intervention itself, and follow-up care to the condition

2. Nonmedical care related to the treatment, such as the cost of traveling to and from the clinic
3. Value of time that family and friends spend caring for the patient
4. Value of the patient's time in treatment

16.3.3.1 Tracking Resources Consumed in Care

Although much of the data on direct medical care can be obtained directly from collection forms designed for the trial itself, it is sometimes necessary to modify the collection forms to collect data that would otherwise not be collected. In cases where clinical forms include only minimal health care utilization data, finding all relevant data can be problematic. One option is to collect billing or insurance records. This option will be viable only if the records are easily obtainable and are comprehensive with regard to capturing all relevant health utilization. Often, this is not the case. For example, Medicare records do not include drug information and some nursing and home care. In cases where billing or insurance records are unobtainable or incomplete, then special health care utilization tracking forms may need to be created. Research coordinators can fill in these forms during interviews at scheduled visits, or they can be filled in periodically by subjects themselves. While no study has compared collection methods directly, researchers must be aware of the potential for recall bias in either form of recording. Short time intervals between collection intervals are necessary to minimize recall problems.¹⁹ For economic data, we recommend monthly data collection.

Tracking nonmedical care resources related to treatment and the value of time lost to treatment, for example, the cost of lost work while seeking medical care, can be especially problematic. Many times, these data are imputed; estimates are made based on interviews with samples of patients or from national databases such as wage rates from the United States Bureau of Labor Statistics. Many researchers choose to explicitly delete nonmedical care from the analyses. In this case, the perspective of the study is no longer societal but rather closer to that of the health insurance payer.

16.3.3.2 Protocol-Induced Costs

Protocol-induced costs are costs incurred for patient care related to the clinical trial that will not be incurred once the intervention is used in practice. There are two types of protocol-induced costs: direct and indirect. Direct costs are costs incurred for tests that are included in a trial for regulatory, scientific, or other reasons beyond clinical care. Direct protocol-induced costs are relatively easy to identify and eliminate from the analysis. Leaders from centers involved in the trial can be polled to identify resources that are unlikely to be utilized outside the clinical trial setting. Indirect protocol-induced costs are costs incurred when investigators respond to abnormalities found during protocol testing if these abnormalities would not have been identified otherwise. For example, if a protocol for a clinical trial of an exercise intervention for emphysema patients calls for preintervention fine resolution computed tomography (CT) scanning of the chest, evaluation costs will occur when scanning identifies pulmonary nodules that would be missed during screening of these patients in

nonexperimental settings. Indirect protocol-induced costs can be quite difficult to identify *ex ante*. If such costs are expected, it is useful to first consider whether there will be differences in these costs between study arms. If not, then there is less concern, because the costs will be eliminated in the incremental analysis (Equation (16.1)). If protocol costs are likely to be different between arms, for example, due to differences in protocol evaluation for each arm, then special measures must be taken to identify them. In extreme cases, chart auditing may be necessary to identify the events and associated resources incurred following the trial.

16.3.3.3 Assigning Values to Resources Consumed

After resources are collected, they must be valued, ideally using nationally representative price weights. When considering prices, it is important to distinguish between charges, reimbursements, and true costs. Charges are the bills that hospitals and health care providers send to payers and patients for health care services. Reimbursement is the amount that is actually paid to the health care providers by patients and the insurer. True costs are what health care providers actually expend to provide the services before markup or profit. Charges, reimbursement levels, and true costs can differ substantially. When the perspective of the study is societal, costs are most appropriate. If an insurer's viewpoint is used, reimbursements are most appropriate. In all cases, using charges to value health resources is inappropriate, because charges almost always greatly overstate the value of the service relative to what it actually costs to provide the service and to what most parties in today's market are willing to pay.

16.4 ANALYSIS OF FINDINGS

16.4.1 ANALYSIS OF COSTS

The analysis of medical care costs raises two methodologically challenging issues. First, analyses are difficult to perform when the proper interval for the cost-effectiveness analysis exceeds the length of the follow-up period in the data for most or all of the study subjects. Second, cost data are difficult to analyze by standard means because of their typically nonstandard statistical distribution. Medical cost distributions often exhibit a mass at zero representing nonusers of medical resources, and skewness representing relatively small numbers of extremely high users. Historically, the methods developed to deal with costs were not designed to accommodate censored data.²⁰ In recent years, nonparametric methods have been described to estimate medical care costs. Nonparametric methods have the advantage of not depending on the statistical distribution of the costs. Lin et al. proposed a method for estimating costs as follows:²¹

$$\hat{C} = \sum_{j=1}^J C_j S_j \quad (16.2)$$

where C_j is the observed mean cost in month j among survivors to month j , S_j is the Kaplan–Meier estimate of survival to month j , and the summation is over months or some other interval after the start of the trial. This estimator uses data from every patient

during each month for which he or she is under observation. For validity, it requires independent censoring in time and representativeness of the observed mean monthly cost. A second nonparametric method by Bang and Tsiatis²² does not pose any restrictions on the distribution of censoring times. The estimator uses cost information from uncensored individuals only and weights these observations by the inverse of the probability of not being censored at the point of death. This estimator is defined as

$$\hat{C} = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i M_i}{K(T_i)} \quad (16.3)$$

where $K(T_i)$ is the probability that the individual i has survived to T_i without being censored, M_i denotes the total cost of the patient during the specified time, and δ takes the value of 1 when the observation is uncensored and 0 otherwise.²² Raikou and McGuire find that both the Lin et al. and Bang and Tsiatis methods produce consistent estimators of average cost; however, the Bang and Tsiatis method becomes less stable at very high levels of censoring.²³ The choice of a method therefore should depend on the degree of censoring observed in the clinical trial.

Stratification is used in RCTs to reduce potential confounding factors that could impact clinical outcomes between the treatment and control groups but may not necessarily address factors that could influence *economic* outcomes. Multivariate methods^{24,25} can be employed to produce adjusted cost estimates in cases where unequal stratification of factors influencing costs is suspected.

16.4.2 PER PROTOCOL VS. INTENT TO TREAT ANALYSES

To preserve the benefits of the randomization process in RCTs, the intent-to-treat approach analyzes individuals in their assigned groups regardless of treatment actually received.²⁶ In most cases, the intent-to-treat design is also appropriate for the economic analysis, but in some cases special circumstances related to the trial or discrepancies between how the intervention is being used in the trial and how it is likely to be used in practice may warrant attention during the design phase. For example, trial investigators for the National Emphysema Treatment Trial, a randomized trial of lung volume reduction surgery vs. medical therapy for patients with severe emphysema, expected that some patients who were randomized to the medical therapy arm would obtain surgery outside the clinical trial.²⁷ Including these patients as originally randomized for analysis under intent-to-treat would bias both the final cost and outcome estimates. In this case, out of protocol use of surgery should be identified and related resource use must be tracked. If the proportion of out of protocol use is small, it is reasonable to include costs and outcomes for those patients as assigned in the trial. If a substantial proportion of patients in the medical therapy arm obtain surgery outside the trial, an additional analysis should be performed excluding such patients and compared to results from the intent-to-treat analysis.

16.4.3 MISSING DATA

Missing data are inevitable in clinical trials and can be handled by a variety of approaches depending on the type of missingness; see Little and Rubin²⁸ and the

chapters by Troxel and Moinpour and Fairclough in this volume. Missing data are also common in economic analyses conducted alongside trials, particularly if researchers for the clinical trial view the extra work of collecting cost or specialized outcome variables as burdensome. In this case, one may find that the degree of missingness for the economic information exceeds that for the clinical information. Recently, researchers have begun to address the issues of missing data in economic analyses alongside clinical trials. The issues that one must consider for missing economic data are not necessarily different than other forms of missing data, with the exception that cost data are often highly skewed. In analyzing datasets with missing data, one must determine the nature of the missing data and then define an approach for dealing with the missing data.

One can address missing data by imputation, replacing the missing field with an estimate of the value. Commonly used methods for imputation include using the mean value from available data, regression analysis using complete variables, *hot deck* imputation, and maximum likelihood approaches. In a recent paper, Briggs et al. outline the advantages and potential problems with these methods for cost data.²⁹ The primary problem with these deterministic methods is that they reduce the variance of the imputed data set. Missing variables are often from severely ill persons on the skewed end of the distribution so that symmetric parametric methods yield inappropriate imputations. Multiple imputation approaches introduce a random component into each imputed value, reflecting the uncertainty that is inherent in missing data.³⁰ Nonparametric bootstrapping is an attractive method for generating such stochastic imputed data for costs. Bayesian simulation methods are also available to address these issues.^{31,32} Most of the commonly used statistical software packages include programs for imputation of missing data. A review of these programs can be found at Multiple Imputation Online.^{33,34}

16.4.4 ADJUSTMENT TO CONSTANT DOLLARS AND DISCOUNTING OF FUTURE COSTS AND OUTCOMES

In cases where costs are collected over a number of years or in several geographic regions, prices must be adjusted to constant nationally representative dollars. Prices can be adjusted to constant dollars for a year of reference using the medical care component of the consumer price index.³⁵

Costs and outcomes stemming from medical interventions should be discounted to a present value to reflect the fact that individuals prefer immediate over delayed monetary gains or health benefits. The present value of future costs can be calculated as follows:

$$C_{\text{present}} = \sum_{j=1}^J \frac{c_j}{(1+r)^{1-j}} \quad (16.4)$$

where C_{present} represents the cost in current dollars, i is the year of the study (year 1 is the first year of the intervention, and year J is the final year of observation), c_j is the cost in year j , and r is the discount rate. Note that the equation is the same when discounting outcomes (e.g., QALYs); simply substitute O_{present} for C_{present} and o_j for c_j .

There is ongoing debate regarding the selection of discount rates for costs and outcomes. Some favor using the same discount rate for health outcomes and costs. Others advocate for differential discounting of costs and outcomes. Some recommend time-varying discount rates, such as 5% over the first 5 years and 3% for later years. The discounting approach chosen can have a substantial impact on the cost-effectiveness estimates of health care programs where most of the costs are incurred in the near future and health benefits occur in the distant future. Uniform discounting of costs and outcomes leads to a prioritizing of treatments with immediate benefit over those with effects that occur in the future, for example, chemotherapy for cancer vs. smoking cessation programs.

Ideally, the discount rate for the costs and health outcomes from health care interventions should be based on actual measures of the social rate of time preference; that is, the rate at which society is willing to trade off present for future consumption. Unfortunately, estimating the actual social time preference is difficult and itself controversial.^{36–38} As a result, the discount rate is usually chosen by panels. In the United States, the U.S. Panel on Cost Effectiveness and Medicine has chosen a 3% discount rate for costs and outcomes.⁵

16.4.5 SUMMARY MEASURES

Equation (16.1) summarizes the results of a CEA as a ratio of the difference of costs to difference in outcomes. It is common to place these results on a cost-effectiveness plane as shown in Figure 16.2.³⁹ Results in the northeast quadrant (higher costs,

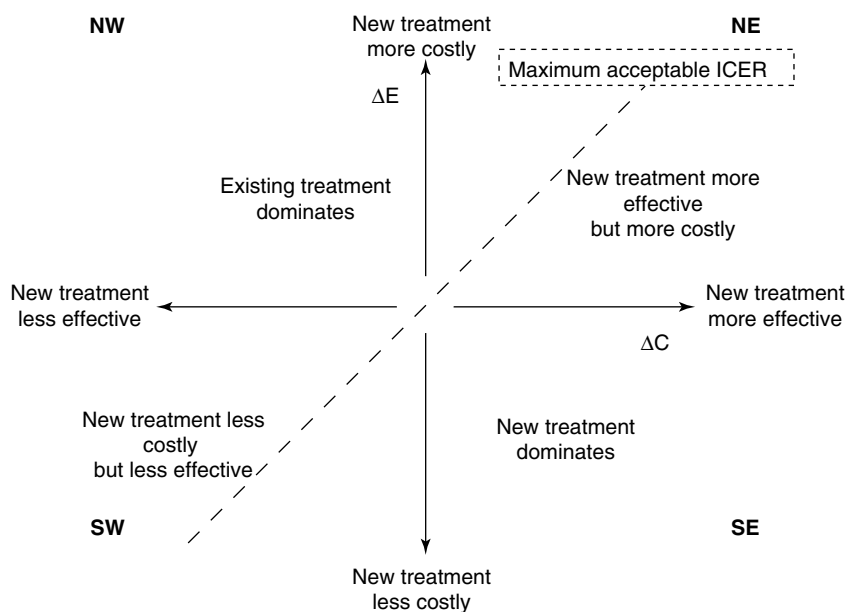


FIGURE 16.2 Cost-effectiveness plane. (Reprinted with permission from Black WC, *Med Decis Making* 1990;10:212–214.)

greater outcomes) are most common. In the positive quadrants, low incremental cost-effectiveness ratios (ICERs) are preferred to high ICERs. Although this is a useful approach to summarizing outcomes, three potential problems can arise for the analyst. The first is when either the difference in costs or effectiveness is negative, such as lower costs or worse outcomes for the intervention group. Unlike situations where the difference in costs and outcomes are both positive, interpreting negative ratios is problematic. In the northwest quadrant, for example, less negative values (smaller difference in effects) are preferable, while in the southeast quadrant more negative values (greater difference in costs) are preferred. This information cannot be easily conveyed, however, when the results are reported as a summary score.

A second potential problem arises when either the cost difference or the effect difference is not statistically significant. In either case, it is common to focus on one dimension of the plane. In the more common case where costs are different but the effect is not, such analyses are called cost-minimization analyses. Briggs and O'Brien have argued that the problem with this simple approach to decision making in situations where either cost or effect is not statistically significant is that it is based on simple and sequential tests of hypotheses.⁴⁰

A third problem with summarizing cost-effectiveness results as a ratio involves evaluation of uncertainty surrounding the ratio. While it is simple to use the bootstrapping technique to create confidence intervals around the ICER, problems of interpretation arise when one end of the confidence bounds crosses zero. As we have discussed above, another possible problem can occur when the difference in effect is very small even when statistically significant. In this case the denominator is unstable, and large or infinite confidence intervals are possible.¹⁰ As will be discussed below, new methods have been developed recently to address the problems inherent in ratio-based summary measures.

16.4.6 COST-EFFECTIVENESS ACCEPTABILITY CURVES

Cost-effectiveness researchers have addressed the problems with ratio based summary measures by replacing the ratio with a decision rule: if the estimated ICER lies below some ceiling ratio λ , reflecting the maximum that decision-makers (or society) are willing to invest to achieve a unit of effectiveness, then the new technology should be implemented. On the cost-effectiveness plane we could summarize uncertainty by determining the proportion of the bootstrap replications that fall below and to the right of a line with slope equal to λ . If there is disagreement about the value of λ , one can report the proportion of bootstrap replications falling below λ as it is varied across a range. Assuming joint normality of cost and effect differences, one can plot the probability that the intervention falls on the cost-effective side of the cost-effectiveness plane as λ varies from 0 to infinity. Such a plot has been termed a cost-effectiveness acceptability curve.⁴¹ Figure 16.3 shows examples of these curves for a cost-effectiveness evaluation of lung volume reduction surgery for severe emphysema.⁴² The curve represents the probability that the intervention is associated with an ICER that is the same or less than the corresponding cost-effectiveness ratios displayed on the x-axis. Note that the median value ($p = 0.5$) corresponds to the base-case ICER. The curve thus allows the analyst to determine the likelihood that

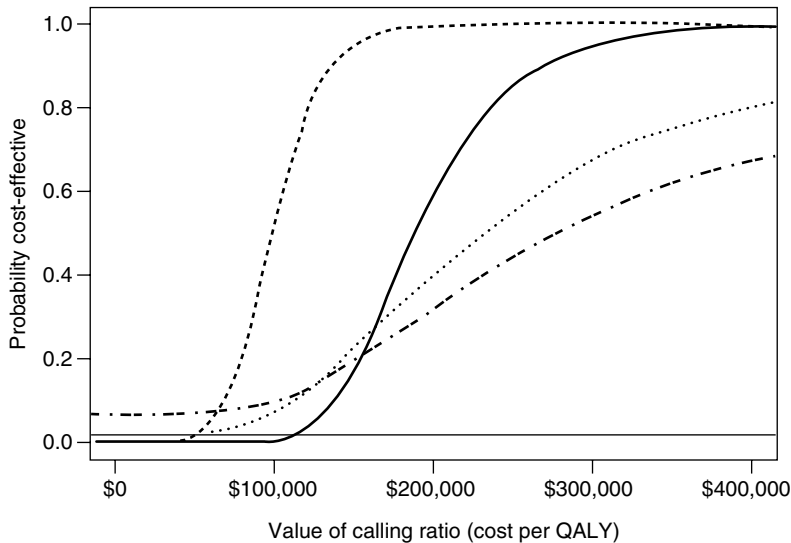


FIGURE 16.3 Cost-effectiveness acceptability curves for lung volume reduction surgery vs. medical therapy for all patients and for three subgroups with significantly improved clinical outcomes in the lung volume reduction surgery arm (either reduced mortality, improved quality of life, or both). The curve represents the probability that lung volume reduction surgery is associated with a cost per quality-adjusted life-year gained that is lower than the corresponding cost-effectiveness ratios displayed on the x-axis. The value of the ceiling ratio at a probability of 0.5 is the median cost per quality adjusted life years for lung volume reduction surgery. Solid horizontal lines denote 95% confidence limits for the projections. (Overall _____; upper-lobe emphysema and low exercise capacity; upper-lobe emphysema and high exercise capacity _____; non-upper-lobe emphysema and low exercise capacity - - - - -.) (Reprinted with permission from Ramsey SD, Berry K, Etzioni R, Kaplan RM, Sullivan SD, Wood DE, National Emphysema Treatment Trial Research Group, *N Engl J Med.* 2003;348:2092–2102.)

the intervention is cost-effective under a wide range of threshold values. The shape of the curve itself also allows the analyst to gauge the level of uncertainty surrounding the results (across all point estimates). Flatter curves denote more uncertainty, while more S-shaped curves suggest less uncertainty surrounding the estimates.

16.4.7 NET HEALTH BENEFIT APPROACH

Another approach analysts have used to address the problem of ratio based summary measures is to eliminate the ratio altogether. Assuming that λ represents a threshold ICER, one can rearrange the equation $ICER = \Delta C/\Delta E \leq \lambda$ as follows:

$$NHB = \Delta E - \frac{\Delta C}{\lambda} \tag{16.5}$$

where NHB is the net health benefit of investing resources in the new technology rather than the established technology. If the $NHB > 0$, the new intervention is cost effective; if $NHB < 0$, more health improvement could be attained by investing resources elsewhere.⁴³ The variance for the NHB can be estimated as:

$$\text{var } NHB = \lambda^2 \text{var}(\Delta E) + \text{var}(\Delta C) = 2\lambda \text{cov}(\Delta E, \Delta C) \quad (16.6)$$

One can compute net benefit statistics as a function of λ , and in this case the interpretation is similar to the cost-effectiveness acceptability curve. In fact, each point of the acceptability curve can be calculated from the p-value of the net-benefit being positive.⁴⁴

REFERENCES

1. Weinstein, M.C. and W.B. Stason, Foundations of cost-effectiveness analysis for health and medical practices. *N Engl J Med*, 1977. 296:716–721.
2. National Institute for Clinical Excellence, *NICE Guidance for Manufacturers and Sponsors*, in Technology Appraisals Process Series 5, June 2001.
3. Fry, R., S. Avey, and S. Sullivan, The Academy of Managed Care Pharmacy Format for Formulary Submissions: a Foundation for Managed Care Pharmacy Task Force Report. *Value in Health*, 2003. 6:505–521.
4. Drummond, M.F., et al., *Methods for the Economic Evaluation of Health Care Programmes*. 2nd ed. 1997, Oxford: Oxford University Press.
5. Weinstein, M.C., et al., Recommendations of the Panel on Cost-effectiveness in Health and Medicine. *Journal of the American Medical Association*, 1996. 276:1253–1258.
6. Adams, M.E., et al., Economic analysis in randomized control trials. *Med Care*, 1992. 30:231–243.
7. Drummond, M.F. and G.L. Stoddart, Economic analysis and clinical trials. *Control Clin Trials*, 1984. 5:115–128.
8. National Institutes of Health, ClinicalTrials.gov, N.L.o. Medicine, Editor. 2004 <<http://www.clinicaltrials.gov/ct>>.
9. O'Brien, B.J., et al., In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Med Care*, 1994. 32:150–63.
10. McIntosh, M., et al., Parameter solicitation for planning cost effectiveness studies with dichotomous outcomes. *Health Econ*, 2001. 10:53–66.
11. Briggs, A.H. and A.M. Gray, Power and sample size calculations for stochastic cost-effectiveness analysis. *Med Decis Making*, 1998. 18:S81–S92.
12. Garber, A.M. and C.E. Phelps, Economic foundations of cost-effectiveness analysis. *J Health Econ*, 1997. 16:1–31.
13. Gold, M.R., et al., eds. *Cost-Effectiveness in Health and Medicine*. 1996, Oxford University Press: New York.
14. Patrick, D.L. and P. Erickson, eds. *Health Status and Health Policy: Quality of Life in Health Care Evaluation and Resource Allocation*. 1993, Oxford University Press: New York.
15. Torrance, G.W., Measurement of health state utilities for economic appraisal. *J Health Econ*, 1986. 5:1–30.

16. Weinstein, M.C., et al., Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices, Modeling Studies. *Value Health*, 2003. 6:9–17.
17. Eddy, D.M., Selecting technologies for assessment. *Int J Technol Assess Health Care*, 1989. 5:485–501.
18. Habbema, J.D., et al., The MISCAN simulation program for the evaluation of screening for disease. *Comput Methods Programs Biomed*, 1985. 20:79–93.
19. Mauskopf, J., et al., A strategy for collecting pharmacoeconomic data during phase II/III clinical trials. *Pharmacoeconomics*, 1996. 9:264–277.
20. Duan, N., Smearing estimate: a nonparametric retransformation method. *J Am Stat Assoc*, 1983. 78:605–610.
21. Lin, D.Y., et al., Estimating medical costs from incomplete follow-up data. *Biometrics*, 1997. 53:419–434.
22. Bang, H. and A.A. Tsiatis, Estimating costs with censored data. *Biometrika*, 2000. 87:329–343.
23. Raikou, M. and A. McGuire, Estimating medical care costs under conditions of censoring. *J Health Econ*, 2004. 23:443–470.
24. Lin, D., Linear regression analysis of censored medical costs. *Biostatistics*, 2000. 1:35–47.
25. Lin, D.Y., Regression analysis of incomplete medical cost data. *Stat Med*, 2003. 22:1181–1200.
26. Fisher, L.D., et al., Intention to treat in clinical trials, in *Statistical Issues in Drug Research and Development*, K.E. Peace, ed. 1990, Dekker: New York. 331–350.
27. Rationale and design of The National Emphysema Treatment Trial: a prospective randomized trial of lung volume reduction surgery. The National Emphysema Treatment Trial Research Group. *Chest*, 1999. 116:1750–1761.
28. Little, R.J.A. and D.B. Rubin, *Statistical Analysis with Missing Data*. 1987, John Wiley & Sons: New York.
29. Briggs, A., et al., Missing ... presumed at random: cost-analysis of incomplete data. *Health Econ*, 2003. 12:377–392.
30. Rubin, D.B., *Multiple Imputation for Nonresponse in Surveys*. 1987, John Wiley: New York.
31. Tanner, M.A. and W.H. Wong, The calculation of posterior distributions by data augmentation. *J Am Stat Assoc*, 1987. 82:528–550.
32. Gelfand, A.E. and A.F.M. Smith, Sampling-based approaches to calculating marginal densities. *J Am Stat Assoc*, 1990. 85:398–409.
33. van Buuren, S., Multiple Imputation Online. 2004, Leiden, The Netherlands: The Department of Statistics of TNO Prevention and Health. <<http://web.inter.nl.net/users/S.van.Buuren/mi/>> Accessed July 23, 2004.
34. Horton, N.J. and S.R. Lipsitz, Multiple imputation in practice: comparison of software packages for regression models with missing variables. *Am Statistician*, 2001. 55:244–254.
35. U.S. Department of Labor and Bureau of Labor Statistics, Consumer Price Indexes. 2004. <<http://www.bls.gov/cpi/>>
36. Feldstein, M.S., The inadequacy of weighted discount rates, in *Cost-Benefit Analysis*, Richard Layard, ed.. 1972, Penguin: London. 140–155.
37. Bradford, D.F., Constraints on government investment opportunities and the choice of discount rate. *Am Econ Rev*, 1975:887–895.
38. Bonneux, L. and E. Birnie, The discount rate in the economic evaluation of prevention: a thought experiment. *J Epidemiol Community Health*, 2001. 55:123–125.

39. Black, W.C., The CE plane: a graphic representation of cost-effectiveness. *Med Decis Making*, 1990. 10:212–214.
40. Briggs, A.H. and B.J. O'Brien, The death of cost-minimization analysis? *Health Econ*, 2001. 10:179–184.
41. van Hout, B.A., et al., Costs, effects, and C/E-ratios alongside a clinical trial. *Health Econ*, 1994. 3:309–319.
42. Ramsey, S.D., et al., Cost effectiveness of lung-volume-reduction surgery for patients with severe emphysema. *N Engl J Med*, 2003. 348:2092–2102.
43. Stinnett, A.A. and J. Mullahy, Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Med Decis Making*, 1998. 18:S68–S80.
44. Briggs, A.H., B.J. O'Brien, and G. Blackhouse, Thinking outside the box: recent advances in the analysis and presentation of uncertainty in cost-effectiveness studies. *Annu Rev Public Health*, 2002. 23:377–401.

Part IV

Exploratory Analyses and Prognostic Factors

17 Prognostic Factor Studies

*Martin Schumacher, Norbert
Holländer, Guido Schwarzer,
and Willi Sauerbrei*

CONTENTS

17.1	Introduction	289
17.2	Design of Prognostic Factor Studies	291
17.3	Examples: Prognostic Studies in Breast Cancer	292
17.3.1	Freiburg DNA study	292
17.3.2	GBSG-2 Study	293
17.4	The Cutpoint Model	295
17.5	Regression Modeling and Related Aspects	299
17.5.1	Assuming a Log-Linear Relationship for Continuous Covariates	299
17.5.2	Categorizing Continuous Covariates	300
17.5.3	Determining Functional Relationships for Continuous Covariates	302
17.5.4	Further Issues	305
17.6	Classification and Regression Trees	306
17.7	Formation and Validation of Risk Groups	309
17.8	Artificial Neural Networks	314
17.9	Assessment of Prognostic Classification Schemes	319
17.10	Sample Size Considerations	322
17.11	Concluding Remarks	325
	Acknowledgment	326
	References	326

17.1 INTRODUCTION

Besides investigations on etiology, epidemiology, and the evaluation of therapies, the identification and assessment of prognostic factors constitutes one of the major tasks in clinical cancer research. Studies on prognostic factors attempt to determine survival probabilities, or more generally, to predict the course of the disease for groups of patients defined by the values of prognostic factors, and to rank the relative importance of various factors. In contrast to therapeutic studies, however, where statistical principles and methods are well developed and generally accepted, this is not the case for the evaluation of prognostic factors. Although some efforts toward an

improvement of this situation have been undertaken,¹⁻⁵ most of the studies investigating prognostic factors are based on historical data lacking precisely defined selection criteria. Furthermore, sample sizes are often far too small to serve as a basis for reliable results. As far as the statistical analysis is concerned, a proper multivariate analysis considering simultaneously the influence of various potential prognostic factors on overall or event-free survival of the patients is not always attempted. Missing values in some or all of the prognostic factors constitute a serious problem that is often underestimated.

In general, the evaluation of prognostic factors based on historical data has the advantages that follow-up and other basic data of patients might be readily available in a database, and that the values of new prognostic factors obtained from stored tissue or blood samples may be added retrospectively. However, such studies are particularly prone to some of the deficiencies mentioned above, including insufficient quality of data on prognostic factors and follow-up data and heterogeneity of the patient population due to different treatment strategies. These issues are often not mentioned in detail in the publication of prognostic studies but might explain at least to some extent why prognostic factors are discussed controversially and why prognostic models derived from such studies are often not accepted for practical use.⁶

There have been some “classical” articles on statistical aspects of prognostic factors in oncology⁷⁻¹¹ that describe the statistical methods and principles that should be used to analyse prognostic factor studies. These articles, however, do not fully address the problem that statistical methods and principles are not adequately applied when analyzing and presenting the results of a prognostic factor study.^{4,6, 12,13} It is therefore a general aim of this chapter not only to present up-dated statistical methodology but also to point out the possible pitfalls when applying these methods to prognostic factor studies. Statistical aspects of prognostic factor studies are also discussed in the monograph on prognostic factors in cancer¹⁴ as well as in some recent textbooks on survival analysis.^{15,16} In order to illustrate important statistical aspects in the evaluation of prognostic factors and to examine the problems associated with such an evaluation in more detail, data from two prognostic factor studies in breast cancer shall serve as illustrative examples. Even before the results of gene expression analyses have been reported on a large-scale basis, the effects of more than 200 potential prognostic factors have been controversially discussed; about 1000 papers have been published in 2001. This illustrates the importance and the unsatisfactory situation in prognostic factors research. A substantial improvement of this situation is possible only with an improvement in the application of statistical methodology and with better reporting (comparable to the CONSORT statement¹⁷ for randomized controlled trials) that provides a more suitable basis for a systematic review of studies for a specific marker of interest. An international working group is currently preparing specific reporting recommendations for prognostic studies.¹⁸ Severe weaknesses of reporting methods and results of prognostic factor studies were recently illustrated by Riley et al.,¹⁹ several resulting difficulties of a systematic summary and a comprehensive assessment of the effect of a prognostic factor are discussed in Altman.²⁰

Throughout this chapter we assume that the reader is familiar with standard statistical methods for survival data to the extent that is presented in more practically

oriented textbooks;^{15,16,21–25} for a deeper understanding of why these methods work we refer to the more theoretically oriented textbooks on survival analysis and counting processes.^{26–28}

17.2 DESIGN OF PROGNOSTIC FACTOR STUDIES

The American Joint Committee on Cancer (AJCC) has established three major criteria for prognostic factors. Factors must be significant, independent, and clinically important.²⁹ According to Gospodarowicz et al.,¹⁴ significant implies that the prognostic factor rarely occurs by chance; independent means that the prognostic factor retains its prognostic value despite the addition of other prognostic factors; and clinically important implies clinical relevance, such as being capable (at least in principle) of influencing patient management and thus outcome.

From these criteria it becomes obvious that statistical aspects will play an important role in the investigation of prognostic factors.^{14, 30–32} That is also emphasized by Simon and Altman,⁴ who give a concise and thoughtful review on statistical aspects of prognostic factor studies in oncology. Recognizing that these will be observational studies, the authors argue that they should be carried out in a way that the same careful design standards are adopted as are used in clinical trials except for randomization. For confirmatory studies that may be seen as comparable to phase III studies in therapeutic research, they listed eleven important requirements that are given in a somewhat shortened version in Table 17.1. From these requirements it can be deduced that prognostic factors should be investigated in carefully planned prospective studies with sufficient numbers of patients and sufficiently long follow-up to observe the endpoint of interest (usually event-free or overall survival). Thus, a prospective observational study where treatment is standardized and everything is planned in advance emerges as the most desirable study design. A slightly different design is represented by a randomized controlled clinical trial where in addition to some therapeutic modalities various prognostic factors are investigated. It is important in such a setting that the prognostic factors of interest are measured either in all

TABLE 17.1

Requirements for Confirmatory Prognostic Factor Studies According to Simon and Altman

-
1. Documentation of intra- and interlaboratory reproducibility of assays
 2. Blinded conduct of laboratory assays
 3. Definition and description of a clear inception cohort
 4. Standardization or randomization of treatment
 5. Detailed statement of hypotheses (in advance)
 6. Justification of sample size based on power calculations
 7. Analysis of additional prognostic value beyond standard prognostic factors
 8. Adjustment of analyses for multiple testing
 9. Avoidance of outcome-orientated cut-off values
 10. Reporting of confidence intervals for effect estimates
 11. Demonstration of subset-specific treatment effects by an appropriate statistical test
-

patients enrolled in the clinical trial or in those patients belonging to a predefined subset. Both designs, however, usually require enormous resources and a long time until results will be available. Thus, a third type of design is used in the vast majority of prognostic factor studies; it can be termed a retrospectively defined historical cohort, where stored tumor tissue or blood samples are available and basic as well as follow-up data on the patients are already documented in a database. To meet the requirements listed in Table 17.1 in such a situation, it is clear that inclusion and exclusion criteria have to be carefully applied. In particular, treatment has to been given in a standardized manner, at least to some sufficient extent. Otherwise, patients for whom these requirements are not fulfilled have to be excluded from the study. If the requirements are followed in a consistent manner, this will usually lead to a drastic reduction in the number of patients eligible for the study as compared to the number of patients originally available in the database. In addition, the follow-up data are often not of such quality as should be the case in a well-conducted clinical trial or prospective study. Thus, if this design is applied, special care is necessary in order to arrive at correct and reproducible results regarding the role of potential prognostic factors.

The types of designs described above also will be represented by the prognostic studies in breast cancer that we will use as illustrative examples and that will be dealt with in more detail in the next section. Other types of designs, for example, nested case-control studies, case-cohort studies or other study types often used in epidemiology,³³ have only been rarely used for the investigation of prognostic factors. Their role and potential use for prognostic factor research has not yet been fully explored. There is one situation where the randomized controlled clinical trial should be the design type of choice: the investigation of so-called predictive factors that indicate whether a specific treatment works in a particular subgroup of patients defined by the predictive factor but not in another subgroup of patients where it may be harmful. Because this is clearly an investigation of treatment-covariate interactions, it should ideally be performed in the setting of a large-scale randomized trial where information on the potential predictive factor is recorded and analyzed by means of appropriate statistical methods.³⁴⁻³⁸

17.3 EXAMPLES: PROGNOSTIC STUDIES IN BREAST CANCER

17.3.1 FREIBURG DNA STUDY

The first study is an observational database consisting of all patients with primary, previously untreated node positive breast cancer who were operated on between 1982 and 1987 in the Department of Gynecology at the University of Freiburg and whose tumor material was available for DNA investigations. Some exclusion criteria were defined retrospectively including history of malignoma, T_4 and/or M_1 tumors according to the TNM classification system of the International Union Against Cancer (UICC),¹⁴ without adjuvant therapy after primary surgery, and older than 80 years. This left 139 patients out of 218 originally investigated for the analysis. This study will be referred to as the Freiburg DNA study.

Eight patient characteristics were investigated. Beside age, number of positive lymph nodes, and size of the primary tumor, the grading score according to Bloom and

Richardson³⁹ as well as estrogen- and progesterone-receptor status were recorded. DNA flow cytometry was used to measure ploidy status of the tumor using a cutpoint of 1.1 for the DNA index and S-phase fraction, which is the percentage of tumor cells in the DNA synthesizing phase obtained by cell cycle analysis. The distribution of these characteristics in the patient population is displayed in Table 17.2(a).

The median follow-up was 83 months. At the time of analysis, 76 events were observed for event-free survival, which was defined as the time from surgery to the first of the following events: occurrence of locoregional recurrence, distant metastasis, second malignancy, or death. Event-free survival was estimated as 50% after five years. Further details of the study can be found in.⁴⁰

17.3.2 GBSG-2 STUDY

The second study is a prospective controlled clinical trial on the treatment of node positive breast cancer patients conducted by the German Breast Cancer Study Group (GBSG); this study will be referred to as the GBSG-2 study.

TABLE 17.2(a)
Patient Characteristics in the Freiburg DNA Breast Cancer Study

Factor	Category	<i>n</i>	(%)
Age	≤50 years	52	(37)
	>50 years	87	(63)
No. of positive lymph nodes	1–3	66	(48)
	4–9	42	(30)
	≥10	31	(22)
Tumor size	≤2 cm	25	(19)
	2–5 cm	73	(54)
	>5 cm	36	(27)
	missing	5	
Tumor grade	1	3	(2)
	2	81	(59)
	3	54	(39)
	missing	1	
Estrogen receptor	≤20 fmol	32	(24)
	>20 fmol	99	(76)
	missing	8	
Progesterone receptor	≤20 fmol	34	(26)
	>20 fmol	98	(74)
	missing	7	
Ploidy status	diploid	61	(44)
	aneuploid	78	(56)
S-phase fraction	<3.1	27	(25)
	3.1–8.4	55	(50)
	>8.4	27	(25)
	missing	30	

The principal eligibility criterion was a histologically verified primary breast cancer of stage T1a-3aN+M0, for example, positive regional lymph nodes but no distant metastases. Primary local treatment was by a modified radical mastectomy (Patey) with en bloc axillary dissection with at least six identifiable lymph nodes. Patients were not older than 65 years of age and presented with a Karnofsky index of at least 60. The study was designed as a comprehensive cohort study.⁴² That is, randomized as well as nonrandomized patients who fulfilled the entry criteria were included and followed according to the study procedures.

The study had a 2×2 factorial design with four adjuvant treatment arms: three versus six cycles of chemotherapy with and without hormonal treatment. Prognostic factors evaluated in the trial were patient's age, menopausal status, tumor size, estrogen and progesterone receptor, tumor grading according to Bloom and Richardson,³⁹ histological tumor type, and number of involved lymph nodes. Histopathologic classification was reexamined, and grading was performed centrally by one reference pathologist for all cases. Event-free survival (EFS) was defined as time from mastectomy to the first occurrence of either locoregional or distant recurrence, contralateral tumor, secondary tumor, or death.

During 6 years, 720 patients were recruited, of whom about two thirds were randomized. Complete data on the seven standard prognostic factors as given in Table 17.2(b) were available for 686 (95.3%) patients, who were taken as the basic patient population for this study. After a median follow-up time of nearly 5 years, 299 events for EFS and 171 deaths were observed. Event-free survival was about 50% at five years. Data from the GBSG-2 study are available from <http://www.blackwellpublishers.co.uk/rss/>.

TABLE 17.2(b)
Patient Characteristics in GBSG-2-Study

Factor	Category	<i>n</i>	(%)
Age	≤45 years	153	(22)
	46–60 years	345	(50)
	>60 years	188	(27)
Menopausal status	pre	290	(42)
	post	396	(58)
Tumor size	≤ 20 mm	180	(26)
	21–30 mm	287	(42)
	>30 mm	219	(32)
Tumor grade	1	81	(12)
	2	444	(65)
	3	161	(24)
No. of positive lymph nodes	1–3	376	(55)
	4–9	207	(30)
	≥ 10	103	(15)
Progesterone receptor	<20 fmol	269	(39)
	≥20 fmol	417	(61)
Estrogen receptor	<20 fmol	262	(38)
	≥20 fmol	424	(62)

17.4 THE CUTPOINT MODEL

In prognostic factor studies, values of the factors considered are often categorized in two or three categories. This may sometimes be done according to medical or biological reasons or may just reflect some consensus in the scientific community. When a new prognostic factor is investigated the choice of such a categorization represented by one or more cutpoints is by no means obvious. Thus, often an attempt is made to derive such cutpoints from the data and to take those cutpoints that give the best separation in the data at hand. In the Freiburg DNA breast cancer study we consider S-phase fraction (SPF) as a new prognostic factor, which it indeed was some years ago.¹² For simplicity, we restrict ourselves to the problem of selecting only one cutpoint and to a so-called univariate analysis. Let Z denote the covariate of interest, in the Freiburg DNA breast cancer data the S-phase fraction as a potential prognostic factor. If this covariate has been measured on a quantitative scale, the proportional hazards⁴³ cutpoint model is defined as

$$\lambda(t | Z > \mu) = \exp(\beta) \lambda(t | Z \leq \mu), \quad t > 0$$

where $\lambda(t | \cdot) = \lim_{h \rightarrow 0} 1/h \Pr(t \leq T < t + h | T \geq t, \cdot)$ denotes the hazard function of the event-free survival time random variable T . The parameter $\theta = \exp(\beta)$ is referred to as the hazard ratio of observations with $Z > \mu$ with respect to observations with $Z \leq \mu$ and is estimated through $\hat{\theta} = \exp(\hat{\beta})$ by maximizing the corresponding partial likelihood⁴³ with given cutpoint μ . The fact that μ is usually unknown makes this a problem of model selection where the cutpoint μ has also to be estimated from the data. A popular approach for this type of data-dependent categorization is the minimum p value method where, within a certain range of the distribution of Z called the selection interval, the cutpoint $\hat{\mu}$ is taken such that the p value for the comparison of observations below and above the cutpoint is a minimum. Applying this method to SPF in the Freiburg DNA breast cancer data, we obtain, based on the logrank test, a cutpoint of $\hat{\mu} = 10.7$ and a minimum p value of $p_{\min} = 0.007$ when using the range between the 10% and the 90% quantile of distribution of Z as the selection interval. Figure 17.1(a) shows the resulting p values as a function of the possible cutpoints considered; Figure 17.1(b) displays the Kaplan–Meier estimates of the event-free survival functions of the groups defined by the estimated cutpoint $\hat{\mu} = 10.7$. The difference in event-free survival looks rather impressive, and the estimated hazard ratio with respect to the dichotomized covariate $I(Z > \hat{\mu})$ using the optimal cutpoint $\hat{\mu} = 10.7$, $\hat{\theta} = 2.37$ is quite large; the corresponding 95% confidence interval is [1.27; 4.44].

Simulating the null hypothesis of no prognostic relevance of SPF with respect to event-free survival ($\beta = 0$), we illustrate that the minimum p value method may lead to a drastic overestimation of the absolute value of the log-hazard ratio.⁴⁴ By a random allocation of the observed values of SPF to the observed survival times, we simulate independence of these two variables, which is equivalent to the null hypothesis $\beta = 0$. This procedure was repeated 100 times, and in each repetition we selected a cutpoint by using the minimum p value method, which is often also referred to as an optimal cutpoint. In the 100 repetitions, we obtained 44 significant ($p_{\min} < 0.05$) results for the logrank test corresponding well to theoretical results outlined in Lausen and Schumacher.⁴⁵

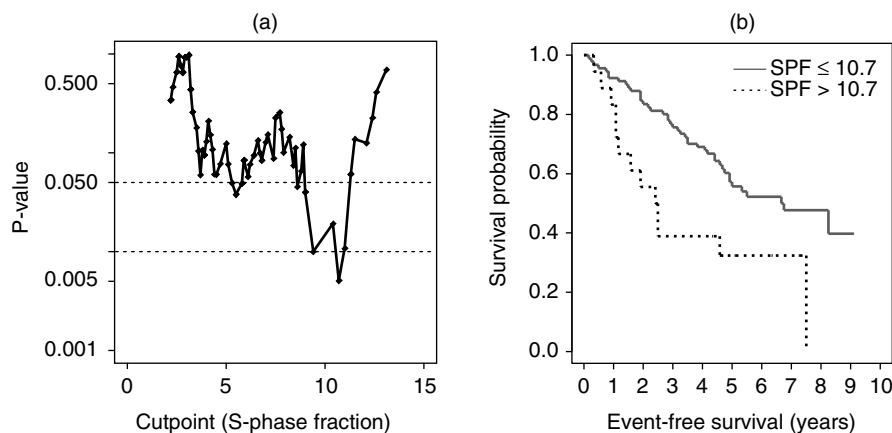


FIGURE 17.1 P -values of the logrank test as a function of all possible cutpoints for S-phase fraction (a) and Kaplan–Meier estimates of event-free survival probabilities by S-phase fraction (b) in the Freiburg DNA study.

The estimated optimal cutpoints of the 100 repetitions and the corresponding estimates of the log-hazard ratio are shown in Figure 17.2(a). We obtained no estimates near the null hypothesis $\beta = 0$ as a result of the optimization process of the minimum p value approach. Due to the well-known problems resulting from multiple testing, it is obvious that the minimum p value method cannot lead to correct results of the logrank test. However, this problem can be solved by using a corrected p value p_{cor} as proposed in Lausen and Schumacher,⁴⁵ developed by taking the minimization process into account. The formula reads

$$p_{\text{cor}} = \varphi(u) \left[u - \frac{1}{u} \right] \log \left[\frac{(1 - \varepsilon)^2}{\varepsilon^2} \right] + 4 \frac{\varphi(u)}{u}$$

where φ denotes the probability density function and u is the $(1 - p_{\text{min}}/2)$ -quantile of the standard normal distribution. The selection interval is characterized by the proportion ε of smallest and largest values of Z that are not considered as potential cutpoints. It should be mentioned that other approaches of correcting the minimum p value could be applied; a comparison of three approaches can be found in a paper by Hilsenbeck and Clark.⁴⁶ In particular, if there are only a few cutpoints, an improved Bonferroni inequality can be applied^{47–49} or the p value correction can be derived from the exact distribution of the maximally selected logrank test.⁵⁰ Using the correction formula given above in the 100 repetitions of our simulation experiment, we obtained five significant results ($p_{\text{cor}} < 0.05$) corresponding well to the significance level of $\alpha = 0.05$. Four significant results were obtained with the usual p value when using the median of the empirical distribution of SPF in the original data as a fixed cutpoint in all repetitions.

In order to correct for overestimation it has been proposed⁵¹ to shrink the parameter estimates by a shrinkage factor c . Considering the cutpoint model the

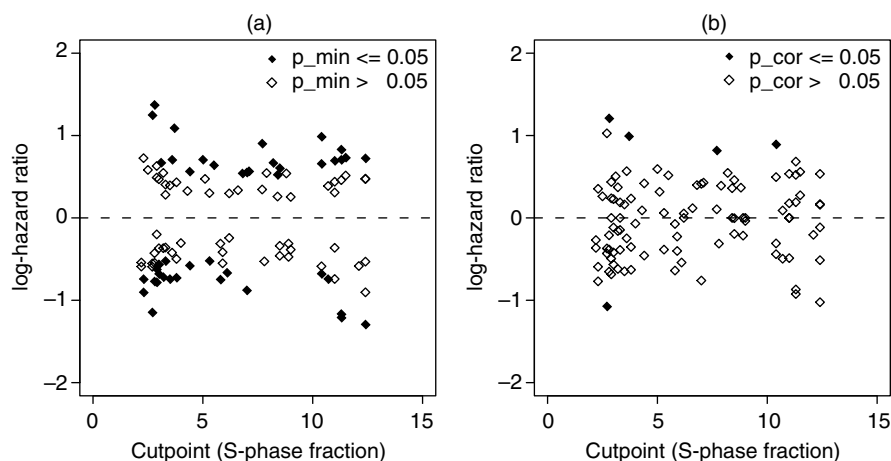


FIGURE 17.2 Estimates of cutpoints and log-hazard ratio in 100 repetitions of randomly allocated observed SPF values to event-free survival times in the Freiburg DNA study before (a) and after (b) correction.

log-hazard ratio should then be estimated by

$$\hat{\beta}_{\text{cor}} = \hat{c} \cdot \hat{\beta}$$

where $\hat{\beta}$ is based on the minimum p value method and $\hat{c}$ is the estimated shrinkage factor. Values of $\hat{c}$ close to 1 should indicate a minor degree of overestimation, whereas small values of $\hat{c}$ should reflect a substantial overestimation of the log-hazard ratio. Obviously, with maximum partial likelihood estimation of c in a model

$$\lambda(t \mid \text{SPF} > \mu) = \exp(c\hat{\beta})\lambda(t \mid \text{SPF} \leq \mu)$$

using the original data, we get $\hat{c} = 1$ because $\hat{\beta}$ is the maximum partial likelihood estimate. In a recent paper⁵² several methods to estimate $\hat{c}$ have been compared. In Figure 17.2(b) the results of the correction process in the 100 simulated studies are displayed when a heuristic estimate $\hat{c} = (\hat{\beta}^2 - \text{var}(\hat{\beta}))/\hat{\beta}^2$ was applied where $\hat{\beta}$ and $\text{var}(\hat{\beta})$ are resulting from the minimum p value method.⁵³ This heuristic estimate performed quite well when compared to more elaborated cross-validation and resampling approaches.⁵²

In general, it has to be recognized that the minimum p value method leads to a dramatic inflation of the type I-error rate. The chance of declaring a quantitative factor as prognostically relevant when in fact it does not have any influence on event-free survival is about 50% when a level of 5% has been intended. Thus, correction of p values is essential but leaves the problem of overestimation of the hazard ratio in absolute terms. The latter problem, which is especially relevant when sample sizes and/or effect sizes are of small to moderate magnitude, could be solved at least partially by applying some shrinkage method. In the Freiburg DNA breast cancer data, we obtain a corrected p value of $p_{\text{cor}} = 0.123$, which provides no clear indication that

S-phase is of prognostic relevance for node-positive breast cancer patients. The correction of the hazard ratio estimate leads to a value of $\hat{\theta}_{\text{cor}} = 2.1$ for the heuristic method and to $\hat{\theta}_{\text{cor}} = 2$ for the cross-validation and bootstrap approaches. Unfortunately, confidence intervals are not straightforward to obtain; bootstrapping the whole model building process including the estimation of a shrinkage factor would be one possibility.⁵⁴

In the 100 repetitions of our simulation experiment, 39 confidence intervals calculated from the model after cutpoint selection did not contain the value $\beta = 0$ (Figure 17.3(a)), corresponding with the number of significant results according to the minimum p value. Although shrinkage is capable of correcting for overestimation of the log-hazard ratio (Figure 17.2(b)), confidence intervals calculated with the estimated model-based standard error do not obtain the desired coverage. In the simulation, there are still 17 out of 100 intervals that do not contain the value $\beta = 0$ (Figure 17.3(b)). Using instead the shrunk risk estimate $\hat{C} \hat{\beta}$ and its empirical variance calculated from 100 bootstrap samples in each repetition of the simulation experiment leads to the correct coverage; only five repetitions do not contain $\beta = 0$. This agrees with the p value correction and corresponds to the significance level of $\alpha = 0.05$. It should be noted, however, that the optimal cutpoint approach still has disadvantages. One of these is that different studies will most likely yield different cutpoints, making comparisons across studies extremely difficult or even impossible. Altman et al.¹² point out this problem for studies of the prognostic relevance of S-phase fraction in breast cancer published in the literature; they identified 19 different cutpoints, some of them motivated only as the optimal cutpoint in a specific dataset. Thus, other approaches, such as regression modeling, might be preferred.

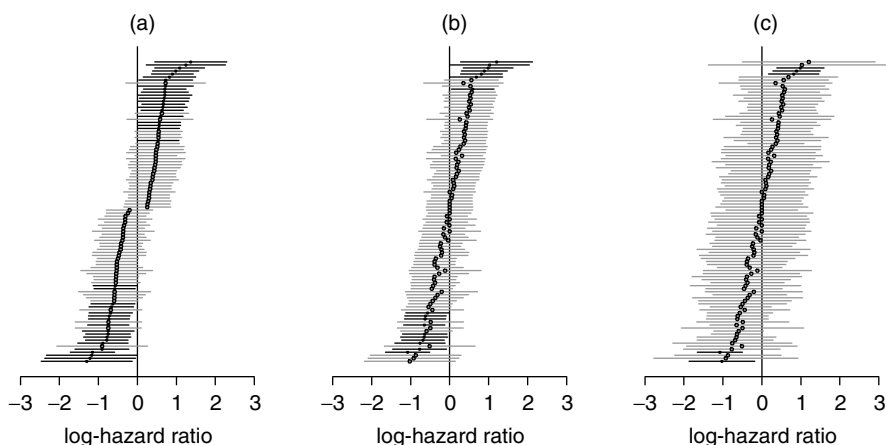


FIGURE 17.3 The 95% confidence intervals for the log-hazard ratio of the S-phase fraction in 100 repetitions of randomly reallocated S-phase values to the observed survival time: estimate based on optimal cutpoints, naïve model-based variance (a); shrunk estimate, naïve model based variance (b); and shrunk estimate, empirical variance from 100 bootstrap samples (c). Confidence intervals not including $\beta = 0$ are marked with filled circles, and the samples are ordered according to the values of $\hat{\beta}$ from smallest to largest.

Taking for example S-phase as a continuous covariate in a Cox regression model, with baseline hazard function $\lambda_0(t)$:

$$\lambda(t|Z) = \lambda_0(t) \exp(\tilde{\beta}Z)$$

yields a p value of $p = 0.061$ for testing the null hypothesis $\tilde{\beta} = 0$.

17.5 REGRESSION MODELING AND RELATED ASPECTS

The Cox proportional hazards regression model^{43,55} is the standard statistical tool to simultaneously analyze multiple prognostic factors. If we denote the prognostic factors under consideration by $Z_1, Z_2, \dots, Z_k$, then the model is given by

$$\lambda(t|Z_1, Z_2, \dots, Z_k) = \lambda_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2 + \dots + \beta_k Z_k)$$

where $\lambda(t|\cdot)$ denotes the hazard function of the event-free or overall survival time random variable T and $\lambda_0(t)$ is the unspecified baseline hazard. The estimated log-hazard ratios $\hat{\beta}_j$ can then be interpreted as estimated effects of the factors Z_j ($j = 1, \dots, k$). If Z_j is measured on a quantitative scale then $\exp(\hat{\beta}_j)$ represents the increase or decrease in risk if Z_j is increased by one unit; if Z_j is a binary covariate then $\exp(\hat{\beta}_j)$ is simply the hazard ratio of category 1 to the reference category ($Z_j = 0$), which is assumed to be constant over the time range considered. It has to be noted that the final multivariate regression model is often the result of a more or less extensive model building process that may involve the categorization and/or transformation of covariates as well as the selection of variables in an automatic or a subjective manner. This model building process should in principle be taken into account when judging the results of a prognostic study; in practice it is often neglected. We will come back to this problem at several occasions below, especially in sections 17.7 and 17.8.

We will demonstrate various approaches with the data of the GBSG-2 study. The factors listed in Table 17.2(b) are investigated with regard to their prognostic relevance. Because all patients received adjuvant chemotherapy in a standardized manner and there appeared no difference between three and six cycles,⁴¹ chemotherapy is not considered any further. Because of the patients' preference in the nonrandomized part and because of a change in the study protocol concerning premenopausal patients, only about a third of the patients received hormonal treatment. Age and menopausal status had a strong influence on whether this therapy was administered. Because hormonal treatment was not of primary interest, all analyses were stratified by it. That is, the baseline hazard was allowed to vary between the two groups defined by hormonal treatment.

17.5.1 ASSUMING A LOG-LINEAR RELATIONSHIP FOR CONTINUOUS COVARIATES

In a first attempt, all quantitative factors are included as continuous covariates. Age is taken in years, tumor size in mm, lymph nodes as number of involved nodes, progesterone and estrogen receptor in fmol/ml. Menopausal status is a binary covariate

coded 0 for premenopausal and 1 for postmenopausal. Grade is considered as a quantitative covariate in this approach, i.e., the risk between grade categories 1 and 2 is the same as between grade categories 2 and 3. The results of this Cox regression model are given in Table 17.3 in terms of estimated hazard ratios and p values of the corresponding Wald tests under the heading “Full model.” In a complete publication, this should at least be accompanied by confidence intervals for the hazard ratios, which we have omitted here for brevity. From this full model it can be seen that tumor size, tumor grade, number of positive lymph nodes, and the progesterone receptor have a significant impact on event-free survival when a significance level of 5% is used. Age, menopausal status and the estrogen receptor do not exhibit prognostic relevance. The full model has the advantage that the regression coefficients of the factors considered can be estimated in an unbiased fashion. However, it is hampered by the facts that the assumed log-linear relationship for quantitative factors may be in sharp contrast to the real situation and that irrelevant factors are included that will not be needed in subsequent steps, for example, in the formation of risk groups defined by the prognostic factors. In addition, correlation between various factors may lead to undesirable statistical properties of the estimated regression coefficients, such as inflation of standard errors or problems of instability caused by multicollinearity. It is therefore desirable to arrive at a simple and parsimonious final model that contains only those prognostic factors that strongly affect event-free survival.⁵⁶ The three other columns of Table 17.3 contain the results of the Cox regression models obtained after backward elimination (BE) for three different selection levels.⁵⁷ A single factor backward elimination with selection level 15.7% (BE (0.157)) corresponds asymptotically to the well-known Akaike information criterion (AIC), whereas selection levels of 5% or even 1% lead to a more stringent selection of factors.⁵⁸ In general, backward elimination can be recommended because of several advantages compared to other stepwise variable selection procedures.^{56,59,60}

In the GBSG-2 study, tumor grade, lymph nodes, and progesterone receptor are selected for all three selection levels considered; when using 15.7% as the selection level, tumor size is included in addition. The AIC or the Bayesian information criterion (BIC),⁶¹ which depends on sample size and puts more penalty on each covariate in the selected model than the AIC, may be used for model assessment. Then the smallest value of AIC or BIC corresponds to the best model.

For the two different selected models, the values of AIC and BIC, respectively, are similar, and at least for AIC the full model seems to be only slightly worse. Thus, the results of the full model and the three backward elimination procedures do not differ too much for these data. However, this should not be expected in general. One reason might be that there is a relatively clear-cut difference between the three strong factors (and perhaps tumor size) and the others that show only a negligible influence on event-free survival in this study.

17.5.2 CATEGORIZING CONTINUOUS COVARIATES

The previous approach implicitly assumes that the influence of a prognostic factor on the hazard function follows a log-linear relationship. By taking lymph nodes as the covariate Z , for example, this means that the risk is increased by the factor $\exp(\hat{\beta})$ if

TABLE 17.3

Estimated Hazard Ratio (HR) and Corresponding p Value in the Cox Regression Models for the GBSC-2 Study. Quantitative prognostic factors are taken as continuous covariates assuming a log-linear relationship

Factor	Full Model		BE (0.157)		BE (0.05)		BE (0.01)	
	HR	<i>p</i> Value	HR	<i>V</i>	HR	<i>p</i> Value	HR	<i>p</i> Value
Age	0.991	0.31	—	—	—	—	—	—
Menopausal status	1.310	0.14	—	—	—	—	—	—
Tumor size	1.008	0.049	1.007	0.061	—	—	—	—
Tumor grade	1.321	0.009	1.325	0.008	1.340	0.006	1.340	0.006
Lymph nodes	1.051	<0.001	1.051	<0.001	1.057	<0.001	1.057	<0.001
Progesterone receptor	0.998	<0.001	0.998	<0.001	0.998	<0.001	0.998	<0.001
Estrogen receptor	1.000	0.67	—	—	—	—	—	—
AIC = $-2\log L + 2p$	3120.18		3116.59		3117.93		3117.93	
BIC = $-2\log L + \log(\tilde{n})p$	3146.09		3131.39		3129.03		3129.03	
<i>p</i> number of covariates in the model, $\tilde{n}$ denotes effective sample size $\tilde{n} = 299$								

the number of positive lymph nodes is increased from m to $m + 1$ for $m = 1, 2$, and so on. This could be a questionable assumption at least for large numbers of positive lymph nodes. For other factors even monotonicity of the log-hazard ratio may be violated, which could result in overlooking an important prognostic factor. Because of this uncertainty the prognostic factors under consideration are often categorized and replaced by dummy variables for the different categories.

In a second analysis of the GBSG-2 study, the categorizations in [Table 17.2\(b\)](#) are used, which were prespecified in accordance to the literature.⁴¹ For those factors with three categories, two binary dummy variables were defined contrasting the corresponding category with the reference category chosen as that with the lowest values. So, for example, lymph nodes were categorized into 1–3, 4–9 and ≥ 10 positive nodes; 1–3 positive nodes serves as the reference category. [Table 17.4](#) displays the results of the Cox regression model for the categorized covariates, where again, the results of the full model are supplemented by those obtained after backward elimination with three selection levels. Elimination of only one dummy variable corresponding to a factor with three categories corresponds to an amalgamation of categories.⁹ In these analyses, tumor grade, lymph nodes, and progesterone receptor again show the strongest effects. Age and menopausal status are also marginally significant and are included in the BE(0.157) model. For age, there is some indication that linearity or even monotonicity of the log-hazard ratio may be violated. Grade categories 2 and 3 do not seem well separated as is suggested by the previous approach presented in [Table 17.3](#) where grade was treated as a continuous covariate; compared to grade 1, the latter approach would lead to estimated hazard ratios of 1.321 and $1.745 = (1.321)^2$ for grades 2 and 3, respectively, in contrast to values of 1.723 and 1.746 when using dummy variables. The use of dummy variables with the coding used here may also be the reason that grade is no longer included by backward elimination with a selection level of 1%. In [Table 17.4](#) we have given the p values of the Wald tests for the two dummy variables separately; alternatively we could also test the two-dimensional vector of corresponding regression coefficients to be zero. In any case this needs two degrees of freedom, whereas when treating grade as a quantitative covariate one degree of freedom would be sufficient. The data of the GBSG-2 study suggest that grade categories 2 and 3 could be amalgamated into one category (grade 2–3); this would lead to an estimated hazard ratio of 1.728 and a corresponding p value of 0.019.

Investigating goodness of fit in terms of AIC and BIC, backward elimination leads to an improvement but differences are rather small.

17.5.3 DETERMINING FUNCTIONAL RELATIONSHIPS FOR CONTINUOUS COVARIATES

The results of the two approaches presented in [Tables 17.3](#) and [17.4](#) show that model building within the framework of a prognostic study has to find a compromise between sufficient flexibility with regard to the functional shape of the underlying log-hazard ratio functions and simplicity of the derived model to avoid problems with serious overfitting and instability. From this point of view, the first approach assuming all relationships to be log-linear may be not flexible enough and may not

TABLE 17.4

Estimated Hazard Ratio (HR) and Corresponding p Values in the Cox Regression Models for the GBSG-2 Study. Prognostic Factors are Categorized as in Table 18.2(b)

Factor		Full Model		BE (0.157)		BE (0.05)		BE (0.01)	
		HR	p Value	HR	p Value	HR	p Value	HR	p Value
Age	≤45	1	—	1	—	—	—	—	—
	45–60	0.672	0.026	0.679	0.030	—	—	—	—
	60	0.687	0.103	0.692	0.108	—	—	—	—
Menopausal status	pre	1	—	1	—	—	—	—	—
	post	1.307	0.120	1.304	0.120	—	—	—	—
Tumor size	≤20	1	—	—	—	—	—	—	—
	21–30	1.240	0.165	—	—	—	—	—	—
	>30	1.316	0.089	—	—	—	—	—	—
Tumor grade	1	1	—	1	—	1	—	—	—
	2	1.723	0.031	1.718	0.032	1.709	0.033	—	—
	3	1.746	0.045	1.783	0.036	1.778	0.037	—	—
Lymph nodes	1–3	1	—	1	—	1	—	1	—
	4–9	1.976	<0.001	2.029	<0.001	2.071	<0.001	2.110	<0.001
	≥10	3.512	<0.001	3.687	<0.001	3.661	<0.001	3.741	<0.001
Progesterone receptor	<20	1	—	1	—	1	—	1	—
	≥20	0.545	<0.001	0.545	<0.001	0.536	<0.001	0.494	<0.001
Estrogen receptor	<20	1	—	—	—	—	—	—	—
	≥20	0.994	0.97	—	—	—	—	—	—
AIC = $-2\log L + 2p$		3087.24		3084.49		3083.58		3085.05	
BIC = $-2\log L + \log(\tilde{n})p$		3128.05		3114.09		3102.09		3096.15	
p number of covariates in the model, $\tilde{n}$ denotes effective sample size $\tilde{n} = 299$									

capture important features of the relationship between various prognostic factors and event-free survival. Conversely, the categorization used in the second approach can always be criticized because of some degree of arbitrariness and subjectivity concerning the number of categories and the specific cutpoints chosen. In addition, it will not fully exploit the information available and will be associated with some loss in efficiency. For a more flexible modeling of the functional relationship, a larger number of cutpoints and corresponding dummy variables would be needed. We will therefore sketch a third approach that will provide more flexibility while preserving simplicity of the final model to an acceptable degree.

The method was originally developed by Royston and Altman⁶² and is termed the fractional polynomial (FP) approach. For a quantitative covariate Z , it uses functions $\beta_0 + \beta_1 Z^p + \beta_2 Z^q$ to model the log-hazard ratio; the powers p and q are taken from the set $\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, and Z^0 is defined as $\log Z$. For practical purposes, the use of two terms is sufficient, and the resulting function is termed a fractional polynomial of degree 2. This simple extension of ordinary polynomials generates a considerable range of curve shapes while still preserving simplicity when compared to smoothing splines or other nonparametric techniques, for example. Furthermore, Holländer and Schumacher⁶³ showed in a simulation study that the data driven selection process, which is used to select the best FP, maintains the type I error rate and generally ends up with a log-linear relationship if it is present. This is in contrast to the optimal cutpoint selection procedure outlined in Section 17.4.

Sauerbrei and Royston⁶⁴ have extended the FP approach, proposing a model building strategy consisting of FP-transformations and selection of variables by backward elimination called the multivariable FP approach. Omitting details of this model building process, which are reported elsewhere,⁶⁴ we have summarized the results in Table 17.5. For age, the powers -2 and -0.5 have been estimated and provide significant contributions to the log-hazard ratio as a function of Z . This function is displayed in Figure 17.4(a) in comparison with the corresponding functions derived from the two other approaches. It provides further indication that there is a nonmonotonic relationship that would be overlooked by the log-linear approach. Grade categories 2 and 3 have been amalgamated as has been pointed out above. For lymph nodes a further restriction has been incorporated by assuming that the relationship

TABLE 17.5
Estimated Regression Coefficients and Corresponding
***p*-Values in the Final Cox Regression Model for the GBSG-2**
Study Using the Fractional Polynomial Approach

Factor/Function	Regression Coefficient	<i>p</i> -Value
(Age/50) ⁻²	1.742	—
(Age/50) ^{-0.5}	-7.812	<0.001*
Tumor grade 1	0	—
Tumor grade 2-3	0.517	0.026
exp (-0.12 × lymph nodes)	-1.981	<0.001
(Progesterone receptor + 1) ^{0.5}	-0.058	<0.001

* *p*-value for both terms of age function

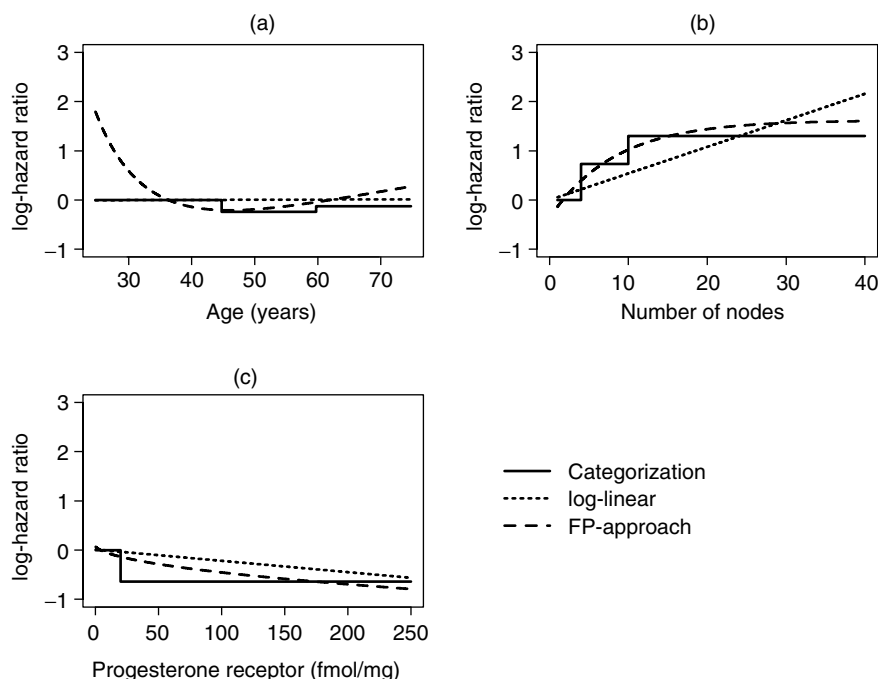


FIGURE 17.4 Estimated log-hazard ratios functions for age (a) lymph nodes (b) and progesterone receptor (c) obtained by the FP, categorization, and log-linear approaches in the GBSG-2 study.

should be monotone with an asymptote for large numbers of positive nodes. This was achieved by using the simple primary transformation $\exp(-0.12 * \text{Lymph nodes})$ where the factor 0.12 was estimated from the data.⁶⁴ The estimated power for this transformed variable was equal to one, and a second power was not needed. Likewise, for progesterone receptor, a power of 0.5 was estimated that gives a significant contribution to the log-hazard ratio functions. For the latter, an improvement in the log-likelihood of about 7.47 was achieved as compared to the inclusion of a linear term only. Details of the process of selecting a specific fractional polynomial function best are illustrated in⁶⁵ for the age function. Figures 17.4(b) and 17.4(c) show these functions for lymph nodes and progesterone receptor, respectively, in comparison with those derived from the log-linear and from the categorization approach. For lymph nodes, it suggests that the log-linear approach underestimates the increase in risk for small numbers of positive nodes, whereas it substantially overestimates it for very large numbers. The categorization approach seems to provide a reasonable compromise for this factor.

17.5.4 FURTHER ISSUES

Some other important issues have not explicitly mentioned so far. One is model checking with regard to the specific assumptions to be made; for this issue, we refer

to textbooks and review articles on survival analysis.^{15,25,66–68} A second issue is concerned with other flexible statistical methods, such as generalized additive models;⁶⁹ a comparison of such methods and the FP approach using the data of the GBSG-2 study is presented in.⁶⁵

Another issue is stability and addresses the question of whether we could replicate the selected final model having different data. Bootstrap resampling has been applied in order to investigate this issue.^{69–72} In each bootstrap sample, the whole model selection or building process is repeated and the results are summarized over the bootstrap samples. We illustrate this procedure for backward elimination with a selection level of 5% in the Cox regression model with quantitative factors included as continuous covariates and assuming a log-linear effect (Table 17.3). In Table 17.6 the inclusion frequencies over 1000 bootstrap samples are given for the prognostic factors under consideration. These frequencies underline that tumor grade, lymph nodes, and progesterone receptor are by far the strongest factors: lymph nodes are always included: progesterone receptor is included in 98%; and tumor grade is included in 62% of the bootstrap samples. The percentage of bootstrap samples where exactly this model — containing these three factors only — is selected is 26.1%. In 60.4% of the bootstrap samples, a model is selected that contains these three factors possibly with other selected factors. These figures might be much lower in other studies where more factors with a weaker effect are investigated. This bootstrap approach has been adapted by Royston and Sauerbrei,⁷³ who investigate the stability of the multivariable fractional polynomial approach. It also provides insight into the interdependencies between different factors or functions selected by inspecting the bivariate or multivariate dependencies between models selected.^{72,73} These investigations underline the nonlinear effect of age on disease-free survival in the GBSG-2 study.

17.6 CLASSIFICATION AND REGRESSION TREES

Hierarchical trees are one approach for nonparametric modeling of the relationship between a response variable and several potential prognostic factors. The book by Breiman et al.⁷⁴ gives a comprehensive description of the method of classification and

TABLE 17.6
Inclusion Frequencies over 1,000 Bootstrap Samples
Using the Backward Elimination Method (BE (0.05)) with
a Selection Level of 5 Percent in the GBSG-2 Study

Factor	Inclusion Frequency
Age	18.2%
Menopausal status	28.8%
Tumor size	38.1%
Tumor grade	62.3%
Lymph nodes	100%
Progesterone receptor	98.1%
Estrogen receptor	8.1%

regression trees (CART) that has been modified and extended in various directions.⁷⁵ We concentrate solely on the application to survival data^{76–80} and will use the abbreviation CART as a synonym for different types of tree-based analyses.

Briefly, the idea of CART is to construct subgroups that are internally as homogeneous as possible with regard to the outcome and externally as separated as possible. Thus, the method leads directly to prognostic subgroups defined by the potential prognostic factors. This is achieved by a recursive tree building algorithm.

As in Section 17.5, we start with k potential prognostic factors $Z_1, Z_2, \dots, Z_k$ that may have an influence on the survival time random variable T . We define a minimum number of patients within a subgroup, $n_{\min}$ say, and prespecify an upper bound for the p values of the logrank test statistic, p_{stop} . Then the tree building algorithm is defined by the following steps.⁴⁸

1. The minimal p value of the logrank statistic is computed for all k factors and all allowable splits within the factors. An allowable split is given by a cutpoint of a quantitative or an ordinal factor within a given range of the distribution of the factor or some bipartition of the classes of a nominal factor.
2. The whole group of patients is split into two subgroups based on the factor and the corresponding cutpoint with the minimal p value, if the minimal p value is smaller or equal to p_{stop} .
3. The partition procedure is stopped if there exists no allowable split, if the minimal p value is greater than p_{stop} , or if the size of the subgroup is smaller than $n_{\min}$.
4. For each of the two resulting subgroups the procedure is repeated.

This tree building algorithm yields a binary tree with a set of patients, a splitting rule, and the minimal p value at each interior node. For patients in the resulting final nodes, various quantities of interest can be computed, such as Kaplan–Meier estimates of event-free survival or hazard ratios with respect to specific references or amalgamated groups.

Because prognostic factors are usually measured on different scales, the number of possible partitions will also be different, leading to the problems that have already been extensively discussed in Section 17.4. Thus, correction of p values and/or restriction to a set of few prespecified cutpoints may be useful to overcome the problem that factors allowing more splits have a higher chance of being selected by the tree building algorithm. Because of multiple testing, the algorithm may be biased in favor of these factors over binary factors with prognostic relevance.

We will illustrate the procedure by means of the GBSG-2 study. If we restrict the possible splits to the range between the 10% and 90% quantiles of the empirical distribution of each factor, then the factor age, for example, will allow 25 splits, whereas the binary factor menopausal status will allow for only one split. Likewise, tumor size will allow for 32 possible splits, tumor grade for only two. Lymph nodes will allow for 10 possible splits; progesterone and estrogen receptors offer 182 and 177 possible cutpoints, respectively. Thus, we decide to use the p value correction as outlined in Section 17.4, and we define $n_{\min} = 20$ and $p_{\text{stop}} = 0.05$. As a splitting criterion we use the logrank test statistic; for simplicity, it is not stratified for hormonal therapy.

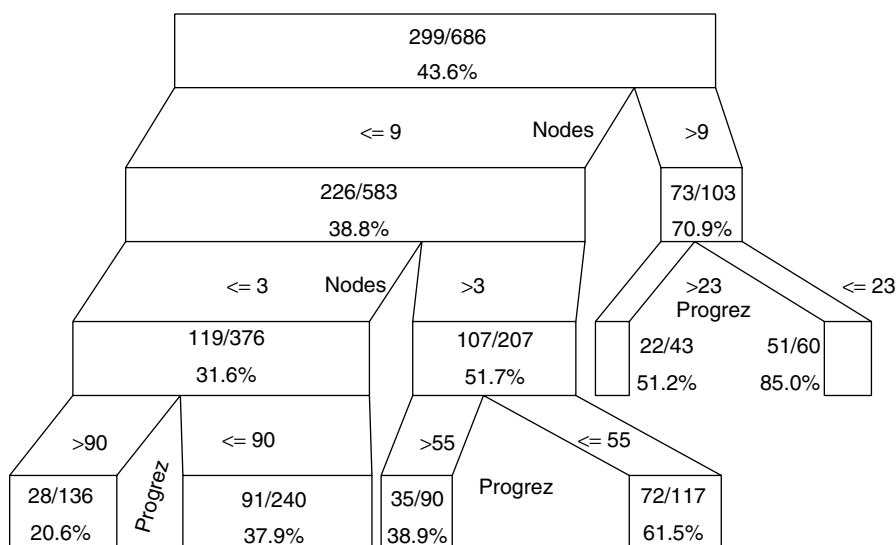


FIGURE 17.5 Classification and regression tree obtained for the GBSG-2 study; P -value correction but no prespecification of cutpoints was used.

The result of the tree-building procedure is summarized in Figure 17.5. In this graphical representation, the size of the subgroups is taken proportional to the width of the boxes, while the centers of the boxes correspond to the observed event rates. This presentation allows an immediate visual impression about the resulting prognostic classification obtained by the final nodes of the tree.

We start with the whole group of 686 patients (the root) where a total of 299 events (event rate 43.6%) has been observed. The factor with the smallest corrected p value is lymph nodes, and the whole group is split at an estimated cutpoint of nine positive nodes ($p_{\text{cor}} < 0.0001$) yielding a subgroup of 583 patients with fewer than or equal to nine positive nodes and a subgroup of 103 patients with more than nine positive nodes. The procedure is then repeated with the left and right nodes. At this level, in the left node lymph nodes again appeared to be the strongest factor, with a cutpoint of three positive nodes ($p_{\text{cor}} < 0.0001$). For the right node, progesterone receptor is associated with the smallest corrected p value and the cutpoint is obtained as 23 fmol ($p_{\text{cor}} = 0.0003$). In these two subgroups, no further splits are possible because of the p_{stop} criterion.

In subgroups of patients with 1 to 3 and 4 to 9 positive nodes, progesterone receptor is again the strongest factor with cutpoints of 90 fmol ($p_{\text{cor}} = 0.006$) and 55 fmol ($p_{\text{cor}} = 0.0018$), respectively. Because of the p_{stop} criterion no further splits are possible.

There exists a variety of definitions of CART-type algorithms that usually consist of tree building, pruning, and amalgamation.^{74,75,81,82} In order to protect against serious overfitting of the data, which in other algorithms is accomplished by tree pruning, we have defined various restrictions, such as the p_{stop} and the n_{min} criteria, and used corrected p values. Applying these restrictions we have obtained the tree displayed in Figure 17.5, which is parsimonious in the sense that only the strongest

factors, lymph nodes and the progesterone receptor, are selected for the splits. However, the values of the cutpoints obtained for progesterone receptor (90, 55, and 23 fmol) are somewhat arbitrary and may not be reproducible or comparable to those obtained in other studies. Another useful restriction may be the definition of a set of prespecified possible cutpoints for each factor. In the GBSG-2 study we specified 35, 40, 45, 50, 55, 60, 65 and 70 years for age; 10, 20, 30, and 40 mm for tumor size; and 5, 10, 20, 100, and 300 fmol for progesterone as well as for estrogen receptor. The resulting tree is displayed in [Figure 17.6\(a\)](#). It only differs from the one without this restriction in the selected cutpoints for the progesterone receptor in the final nodes. For comparison, trees without the p value correction and with and without prespecification of a set of possible cutpoints are presented in [Figures 17.6\(b\) and 17.6\(c\)](#). Because lymph nodes and progesterone receptor are the dominating prognostic factors in this patient population, the resulting trees are identical at the first two levels to those where the p values have been corrected. The final nodes, however, are again split, leading to a larger number of final nodes. In addition, other factors like age, tumor size, and estrogen receptor are now used for the splits at subsequent nodes. A more detailed investigation of the influence of p value correction and prespecification of possible cutpoints on resulting trees and their stability is given by Sauerbrei.⁸³

To improve the predictive ability of trees, stabilizing methods based on resampling, such as bagging have been proposed.^{84–88} However, interpretation of the results is difficult, which limits their value for practical applications. For the future it is important to learn more about the advantages and disadvantages of statistical vs. machine learning methods.⁸⁹

17.7 FORMATION AND VALIDATION OF RISK GROUPS

While the final nodes of a regression tree define a prognostic classification scheme, some combination of final nodes to a prognostic subgroup might be indicated. This is especially important if the number of final nodes is large and/or if the prognosis of patients in different final nodes is comparable. So, for example, from the regression tree presented in [Figure 17.6\(a\)](#), the simple prognostic classification given in [Table 17.7](#) can be derived, which is broadly in agreement to current knowledge about the prognosis of node-positive breast cancer patients. The results in terms of estimated event-free survival are displayed in [Figure 17.7\(a\)](#); the Kaplan–Meier curves show a good separation of the four prognostic subgroups. Because in other studies or in clinical practice progesterone receptor may often be only recorded as positive or negative, the prognostic classification scheme in [Table 17.7](#) may be modified in the way that the definition of subgroups I and II are replaced by

$$\text{I}^*: (\text{LN} \leq 3 \text{ and } \text{PR} > 20)$$

and

$$\text{II}^*: (\text{LN} \leq 3 \text{ and } \text{PR} \leq 20) \text{ or } (\text{LN } 4\text{--}9 \text{ and } \text{PR} > 20)$$

respectively, where LN is lymph nodes and PR is progesterone receptor. The resulting Kaplan–Meier estimates of event-free survival are depicted in [Figure 17.7\(b\)](#).

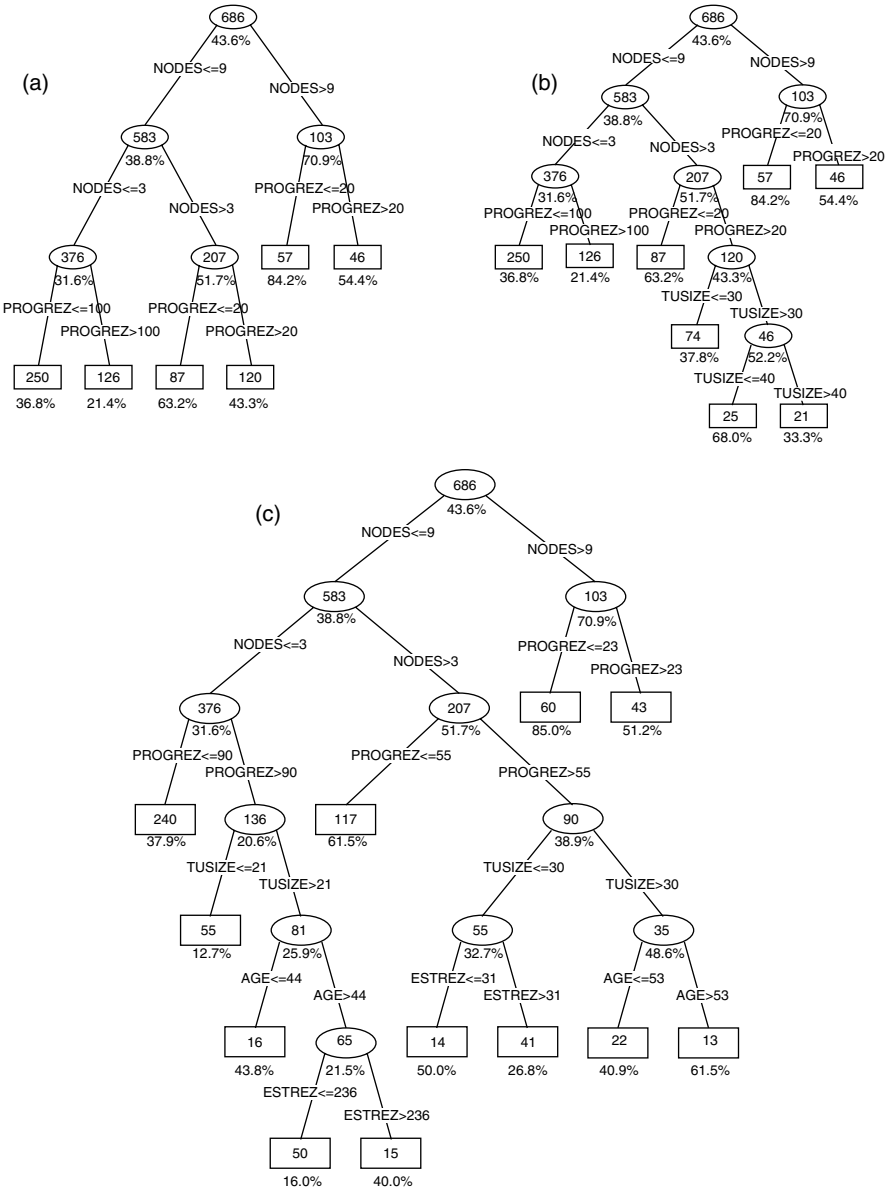


FIGURE 17.6 Classification and regression trees obtained for the GBSG-2 study (P -value correction and prespecification of cutpoints (a); no P -value correction with (b) and without (c) prespecification of cutpoints).

For two of the regression approaches in Section 17.5, prognostic subgroups have been formed by dividing the distribution of the prognostic index, $\hat{\beta}_1 Z_1 + \dots + \hat{\beta}_k Z_k$, into quartiles; the stratified event-free survival curves are displayed in [Figures 17.8\(a\)](#) (Cox regression model with continuous factors, BE(0.05), [Table 17.3](#)) and

TABLE 17.7
Prognostic Classification Scheme Derived from the Regression Tree (*p*-Value Correction and Predefined Cutpoints) in the GBSG-2 Study

Prognostic Subgroup	Definition of Subgroup (LN: No. of Positive Lymph nodes; PR: Progesterone Receptor)
I	(LN≤3 and PR>100)
II	(LN≤3 and PR≤100) or (LN 4–9 and PR>20)
III	(LN 4–9 and PR≤20) or (LN>9 and PR>20)
IV	(LN>9 and PR≤20)

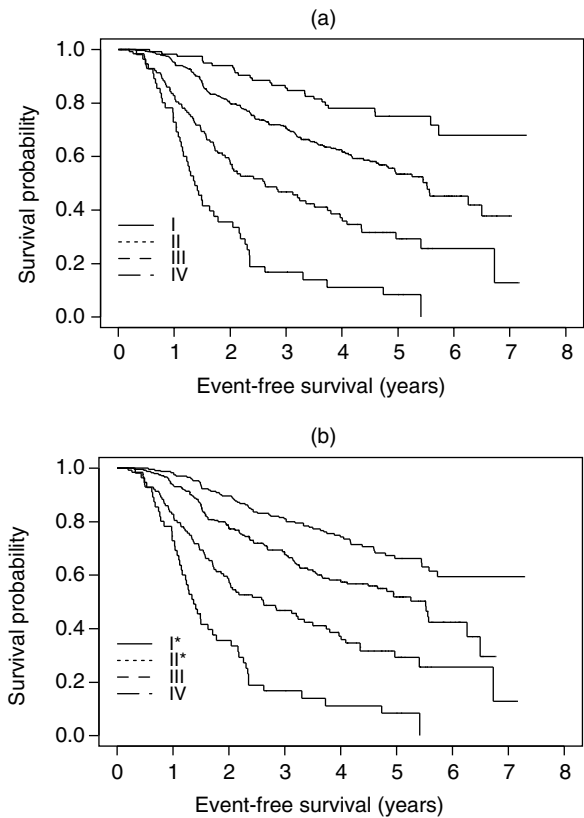


FIGURE 17.7 Kaplan–Meier estimates of event-free survival probabilities for the prognostic subgroups derived from the CART-approach (a) and the modified CART-approach (b) in the GBSG-2-study.

17.8(b) (Cox regression model with categorized covariates, BE(0.05), [Table 17.4](#)). As a reminder, in the definition of the corresponding subgroups tumor grade also enters in addition to lymph nodes and progesterone receptor.

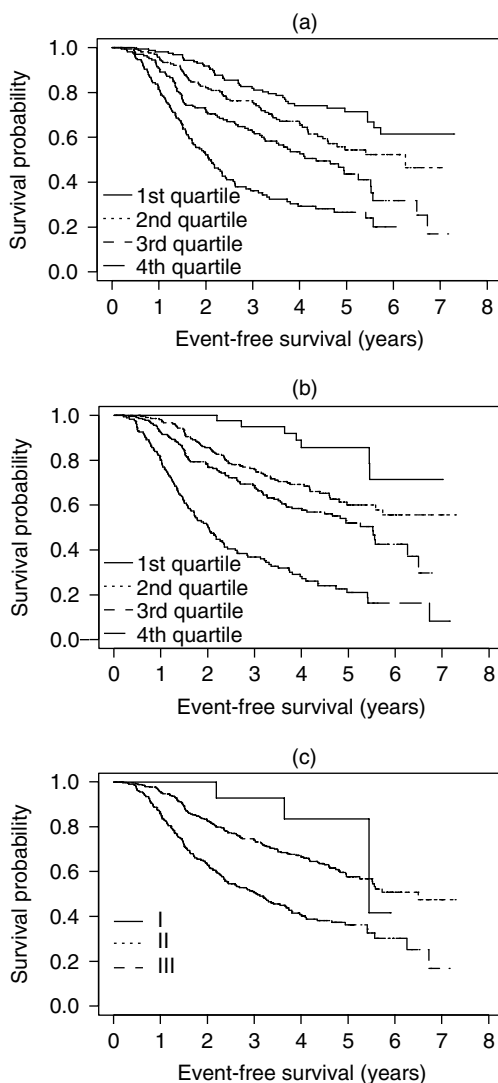


FIGURE 17.8 Kaplan–Meier estimates of event-free survival probabilities for the prognostic subgroups derived from a Cox model with continuous (a) and categorized (b) covariates and according to the Nottingham Prognostic Index (c) in the GBSG-2 study.

For comparison, Figure 17.8(c) shows the Kaplan–Meier estimates of event-free survival for the well-known Nottingham Prognostic Index,^{90,91} the only prognostic classification scheme based on standard prognostic factors that enjoys widespread acceptance.⁹² This index is defined as

$$\text{NPI} = 0.02 \times \text{size (in mm)} + \text{lymph node stage} + \text{tumor grade}$$

where lymph node stage is equal to 1 for node-negative patients, 2 for patients with 1 to 3 positive lymph nodes, and 3 if four or more lymph nodes were involved. It is

usually divided into three prognostic subgroups NPI-I ($\text{NPI} < 3.4$), NPI-II ($3.4 \leq \text{NPI} \leq 5.4$), and NPI-III ($\text{NPI} > 5.4$). Because it was developed for node-negative and node-positive patients, there seems room for improvement by taking other factors, such as progesterone receptor, into account.⁹³

Because the Nottingham Prognostic Index has been validated in various other studies,⁹² we can argue that the degree of separation that is displayed in Figure 17.8(c) could be achieved in general. This, however, is by no means true for the other proposals derived by regression modeling or CART techniques, where some shrinkage has to be expected.^{53,56,94,95} We therefore attempted to validate the prognostic classification schemes defined above with the data of an independent study that in more technical terms is often referred to as a test set.⁹⁶ As a test set, we take the Freiburg DNA study that covers the same patient population and prognostic factors as in the GBSG-2 study. Only progesterone and estrogen receptor status (positive: > 20 fmol and negative: ≤ 20 fmol) is recorded in the Freiburg DNA study, and the original values are not available. Thus, only those classification schemes where progesterone receptor enters as positive or negative can be considered for validation. Furthermore, we restrict ourselves to those patients where the required information on prognostic factors is complete. Table 17.8(a) shows the estimated hazard ratios for the prognostic groups derived from the categorized Cox model and from the modified CART classifications scheme defined above. The hazard ratios have been estimated by using dummy variables defining the risk groups and by taking the group with the best prognosis as reference. When applying the classification schemes to the data of the Freiburg DNA study, the definitions and categorization derived in the GBSG-2 study are used. Note that the categorization into quartiles of the prognostic index does not yield groups with equal numbers of patients because the prognostic index from the categorized Cox model takes only few different values.

From the values given in Table 17.8(a), it can be seen that there is some shrinkage in the hazard ratios when estimated in the Freiburg DNA test set. This shrinkage

TABLE 17.8(a)
Estimated Hazard Ratios for Various Prognostic Classification Schemes Derived in the GBSG-2 Study and Validated in the Freiburg DNA Study

Prognostic Groups		Estimated Hazard Ratios (No. of Patients)			
		GBSG-2 Study		Freiburg DNA Study	
Cox	I	1	(52)	1	(33)
	II	2.68	(218)	1.78	(26)
	III	3.95	(236)	3.52	(58)
	IV	9.92	(180)	7.13	(14)
CART [†]	I*	1	(243)	1	(50)
	II*	1.82	(253)	1.99	(38)
	III	3.48	(133)	3.19	(33)
	IV	8.20	(57)	4.34	(11)
NPI	II	1	(367)	1	(46)
	III	2.15	(301)	2.91	(87)

[†]The subgroups I* and II* are defined on page 309.

is more pronounced in the modified CART classification scheme (reduction by 47% in the relatively small high risk group) as compared with the categorized Cox model (reduction by 28% in the high risk group).

In order to get some idea of the amount of shrinkage that has to be anticipated in a test set, based on the training set⁹⁶ where the classification scheme has been developed, cross-validation or other resampling methods can be used. Similar techniques as in Section 18.4 can be used, which essentially estimate a shrinkage factor for the prognostic index.^{51,53} The hazard ratios for the prognostic subgroups are then estimated by categorizing the shrunk prognostic index according to the cutpoints used in the original data. In the GBSG-2 study we obtained an estimated shrinkage factor of $\hat{c} = 0.95$ for the prognostic index derived from the categorized Cox model indicating that we would not expect a serious shrinkage of the hazard ratios between the prognostic subgroups. Compared to the estimated hazard ratios in the Freiburg DNA study (Table 17.8(a)), it is clear that the shrinkage effect in the test set can be predicted only to a limited extent. This deserves at least two comments. First, we have used leave-one-out cross-validation that possibly could be improved by bootstrap or other resampling methods;⁵² second, we did not take the variable selection process into account. By doing so, we would expect more realistic estimates of the shrinkage effect in an independent study. Similar techniques can in principle also be applied to classification schemes derived by CART methods. How to do this best, however, is still a matter of ongoing research. Further approaches to assess the predictive ability of prognostic classification schemes will be presented in Section 17.9.

17.8 ARTIFICIAL NEURAL NETWORKS

The application of artificial neural networks (ANNs) for prognostic and diagnostic classification in clinical medicine has attracted growing interest in the medical literature. So, for example, a mini-series on neural networks that appeared in the *Lancet* contained three more-or-less enthusiastic review articles⁹⁷⁻⁹⁹ and an additional commentary expressing some scepticism.¹⁰⁰ In particular, feed-forward neural networks have been used extensively, often accompanied by exaggerated statements of their potential. In a recent review article,¹⁰¹ we identified a substantial number of articles with application of ANNs to prognostic classification in oncology.

The relationship between ANNs and statistical methods, especially logistic regression models, has been described in several articles.¹⁰²⁻¹⁰⁷ Briefly, the conditional probability that a binary outcome variable Y is equal to 1, given the values of k prognostic factors $Z_1, Z_2, \dots, Z_k$ is given by a function $f(Z, w)$. In feed-forward neural networks this function is defined by $f(Z, w) = \Lambda\left(W_0 + \sum_{j=1}^r W_j \cdot \Lambda\left(w_{0j} + \sum_{i=1}^k w_{ij} Z_i\right)\right)$ where $w = (W_0, \dots, W_r, w_{01}, \dots, w_{kr})$ are unknown parameters called weights and $\Lambda(\cdot)$ denotes the logistic function, $\Lambda(u) = (1 + \exp(-u))^{-1}$, called activation-function. The weights w can be estimated from the data via maximum likelihood although other optimization procedures are often used in this framework. The ANN is usually introduced by a graphical representation like that in Figure 17.9. This figure illustrates a feed-forward neural network with one hidden layer. The network consists of k input units, r hidden units denoted by $h_0, \dots, h_r$, and one output unit and corresponds to the ANN with $f(Z, w)$ defined above. The arrows indicate the flow of

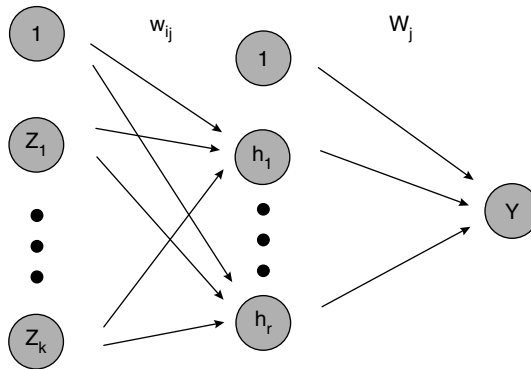


FIGURE 17.9 Graphical representation of an artificial neural network with one input, one hidden, and one output layer.

information. The number of weights is $(r + 1) + (k + 1)r$ because every input unit is connected with every hidden unit and the latter are all connected to the output unit. If there is no hidden layer ($r = 0$), the ANN reduces to a common logistic regression model, which is also called the logistic perceptron.

In general, feed-forward neural networks with one hidden layer are universal approximators¹⁰⁸ and thus can approximate any function defined by the conditional probability that Y is equal to one given Z with arbitrary precision by increasing the number of hidden units. This flexibility can lead to serious overfitting, which can again be compensated for by introducing some weight decay,^{96,109} for example, by adding a penalty term

$$-\lambda \left(\sum_{j=1}^r W_j^2 + \sum_{j=1}^r \sum_{i=1}^k w_{ij}^2 \right)$$

to the log-likelihood. The smoothness of the resulting function is then controlled by the decay parameter λ . It is interesting to note that in our literature review of articles published between 1991 and 1995 we have not found any application in oncology where weight decay has been used.¹⁰¹

Extension to survival data with censored observations is associated with various problems. Although there is a relatively straightforward extension of ANNs to handle grouped survival data¹¹⁰ several naive proposals can be found in the literature. In order to predict outcome (death or recurrence) of individual breast cancer patients, Ravdin and Clark¹¹¹ and Ravdin et al.¹¹² use a network with only one output unit but using the number j of the time interval as additional input. Moreover, they consider the unconditional probability of dying before t_j rather than the conditional one as output. Their underlying model then reads

$$P(T < t_j | Z) = \Lambda \left(w_0 + \sum_{i=1}^k w_i Z_i + w_{k+1} \cdot j \right)$$

for $j = 1, \dots, J$. T denotes again the survival time random variable, and the time intervals are defined through $t_{j-1} \leq t < t_j$, $0 = t_0 < t_1 < \dots < t_J < \infty$. This parameterization

ensures monotonicity of the survival probabilities but also implies a rather stringent and unusual shape of the survival distribution, because in the case of no covariates this reduces to

$$P(T < t_j) = \Lambda(w_0 + w_{k+1} \cdot j)$$

for $j = 1, \dots, J$. Obviously, the survival probabilities do not depend on the length of the time intervals, which is a rather strange and undesirable feature. Including a hidden layer in this expression is a straightforward extension retaining all the features summarized above. De Laurentiis and Ravdin¹¹³ call this type of neural networks time coded models. Another form of neural networks that has been applied to survival data are the so-called single time point models.¹¹³ Because they are identical to a logistic perceptron or a feed-forward neural network with a hidden layer, they correspond to fits of logistic regression models or their generalizations to survival data. In practice, a single time point t^* is fixed and the network is trained to predict the survival probability. The corresponding model is given by

$$P(T < t^* | Z) = \Lambda\left(w_0 + \sum_{i=1}^k w_i Z_i\right)$$

or its generalization when introducing a hidden layer. This approach is used by Burke¹¹⁴ to predict 10-year survival of breast cancer patients based on various patient and tumor characteristics at time of primary diagnosis. McGuire et al.¹¹⁵ utilized this approach to predict 5-year event-free survival of patients with axillary node-negative breast cancer based on seven potentially prognostic variables. Kappen and Neijt¹¹⁶ used it to predict 2-year survival of patients with advanced ovarian cancer obtained from 17 pretreatment characteristics. The neural network they actually used reduced to a logistic perceptron.

The procedure above can be repeatedly applied for the prediction of survival probabilities at fixed time points $t^* = t_1 < t_2 < \dots < t_j$, replacing w_0 by w_{0j} and w_i by w_{ij} for $j = 1, \dots, J$.¹¹⁶ But without restriction on the parameters such an approach does not guarantee that the probabilities $P(T < t_j | Z)$ increase with j , and hence may result in life-table estimators suggesting a nonmonotone survival function. Closely related are multiple time point models¹¹³ where one neural network with J output units with or without a hidden layer is used.

A common drawback of these naive approaches is that they do not allow one to incorporate censored observations in a straightforward manner, which is closely related to the fact that they are based on unconditional survival probabilities instead of conditional survival probabilities. Neither omission of the censored observations, as suggested by Burke,¹¹⁴ nor treating censored observations as uncensored are valid approaches, but both instead are a serious source of bias, which is well known in the statistical literature. De Laurentiis and Ravdin¹¹³ and Ripley¹⁰⁹ propose to impute estimated conditional survival probabilities for the censored cases from a Cox regression model.

Faraggi and Simon¹¹⁷ and others^{118–121} proposed a neural network generalization of the Cox regression model defined by

$$\lambda(t | Z_1, \dots, Z_k) = \lambda_0(t) \exp(f_{FS}(Z, w))$$

where

$$f_{FS}(Z, w) = \sum_{j=1}^r W_j \Lambda \left(w_{0j} + \sum_{i=1}^k w_{ij} Z_i \right)$$

Note that the constant W_0 is omitted in the framework of the Cox model. Estimation of weights is then done by maximizing the partial likelihood. Although the problem of censoring is satisfactorily solved in this approach, there remain problems with potentially serious overfitting of the data, especially if the number r of hidden units is large.

For illustration, we consider factors included in the final FP model of the GBSG-2 study (Section 17.5, Table 17.5). Thus, we used the four factors, age, grade, lymph nodes, and progesterone receptor (all scaled to the interval $[0; 1]$), and hormone therapy as inputs for the Faraggi and Simon network. Figure 17.10 shows the results for various Faraggi and Simon networks compared to the FP approach in terms of Kaplan–Meier estimates of event-free survival in the prognostic subgroups defined by the quartiles of the corresponding prognostic indices. It should be noted that the Faraggi and Simon network contains $5 + (6 \times 5) = 35$ parameters when 5 hidden units are used and $20 + (6 \times 20) = 140$ when 20 hidden units are used. The latter must be suspected of serious overfitting with a high chance that the degree of separation achieved could never be reproduced in other studies. In order to highlight this phenomenon we trained a slightly different Faraggi and Simon (F&S) network where, in addition to age, tumor size, tumor grade, and number of lymph nodes,

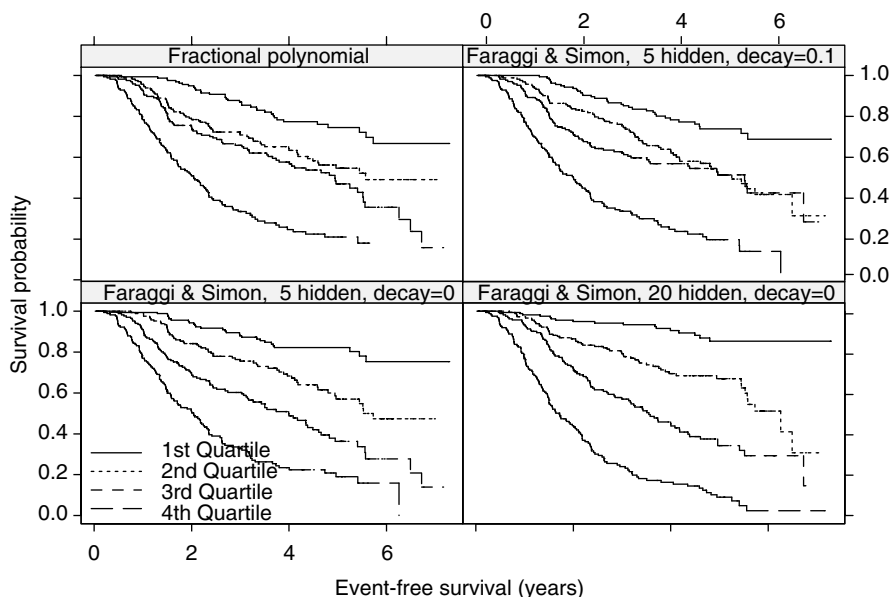


FIGURE 17.10 Kaplan–Meier estimates of event-free survival probabilities for the prognostic subgroups derived from various Faraggi and Simon networks and from the FP approach in the GBSG-2 study.

estrogen and progesterone status (positive: >20 fmol, negative: ≤ 20 fmol) were used as inputs. This network contained 20 hidden units ($20 + (7 \times 20) = 160$ parameters) and showed a similar separation to the one where estrogen and progesterone receptor entered as quantitative inputs. Table 17.8(b) contrasts the results from the GBSG-2 study used as training set and the Freiburg DNA study used as test set in terms of estimated hazard ratios, where the predicted event-free survival probabilities are categorized in quartiles. In the training set, we observe a twenty-fold increase in risk between the high-risk and the low-risk group. But in the test set, the F&S network turns out to yield a completely useless prognostic classification scheme; the estimated hazard ratios are not even monotone increasing. It is obvious that some restrictions, either in terms of a maximum number of parameters or a weight decay, are necessary to avoid such overfitting, especially in the two F&S networks where weight decay was not applied. The results for a network with 5 hidden units are comparable to the FP approach, especially when some weight decay is introduced. It should be noted that the FP approach contains at most 8 parameters if we ignore the preselection of the four factors.

TABLE 17.8(b)

Estimated Hazard Ratios for Various Prognostic Classification Schemes Based on Faraggi and Simon (F&S) Neural Networks Derived in the GBSG-2 Study and Validated in the Freiburg DNA Study

Prognostic Groups		Estimated Hazard Ratios (No. of Patients)			
		GBSG-2 Study		Freiburg DNA Study	
F&S ^a	I	1	(179)	1	(37)
	II	3.24	(178)	0.34	(16)
	III	7.00	(159)	0.98	(38)
	IV	22.03	(170)	1.39	(35)
F&S ^b	I	1	(171)	1	(23)
	II	1.45	(172)	1.57	(25)
	III	2.62	(171)	3.09	(32)
	IV	5.75	(172)	4.27	(46)
F&S ^c	I	1	(172)	1	(20)
	II	1.64	(171)	1.03	(31)
	III	3.27	(171)	1.89	(28)
	IV	8.49	(172)	2.72	(47)
F&S ^d	I	1	(172)	1	(23)
	II	1.46	(171)	1.57	(25)
	III	2.62	(171)	3.22	(33)
	IV	5.77	(172)	4.14	(45)

^a 20 hidden units, weight decay = 0

^b 20 hidden units, weight decay = 0.1

^c 5 hidden units, weight decay = 0

^d 5 hidden units, weight decay = 0.1

17.9 ASSESSMENT OF PROGNOSTIC CLASSIFICATION SCHEMES

Once a prognostic classification scheme is developed, the question arises as to how its predictive ability can be assessed and compared to competitors. There is no commonly agreed upon approach, and most measures are somewhat ad hoc. Suppose that a prognostic classification scheme consists of g prognostic groups, called risk strata or risk groups; then one common approach is to present the Kaplan–Meier estimates for event-free or overall survival in the g groups. This is the way we presented the results of prognostic classification schemes in previous sections. The resulting figures are often accompanied by p values of the logrank test for the null hypothesis that the survival functions in the g risk strata are equal. It is clear that a significant result is a necessary but not sufficient condition for good predictive ability. Sometimes, a Cox model using dummy variates for the risk strata is fitted, and the log-likelihood and/or estimated hazard ratios of risk strata with respect to a reference are given. Recently, we have proposed a summary measure of separation¹²² defined as

$$\text{SEP} = \exp \left[\sum_{j=1}^g \frac{n_j}{n} |\hat{\beta}_j| \right]$$

where n_j denotes the number of patients in risk stratum j , and $\hat{\beta}_j$ is the estimated log-hazard ratio or log-relative risk of patients in risk stratum j with respect to a baseline reference. Models are favored that have large SEP values. Although SEP was designed for use with survival data, it is also applicable to binary data. Algebraically, for survival data SEP turns out to be essentially an estimate of the standard deviation of the predicted log hazard ratios according to a model with a dummy variable for each group. SEP is fairly independent of the number of groups employed. It has the advantage that it can be calculated for models that generate only groups and no risk score, such as tree-based methods (CART), simple schemes based on counting adverse risk factors or expert subjective opinion. SEP has some drawbacks that motivated Royston and Sauerbrei¹²³ to propose improved measures of separation taking the risk ordering across individual patients into account.

We now briefly outline another recent development complementing the measure of separation; a detailed description can be found elsewhere.^{124,125} First, it is important to recognize that the time-to-event itself cannot adequately be predicted.^{126–130} The best one can do at $t = 0$ is to try to estimate the probability that the event of interest will not occur until a prespecified time horizon represented by some time point t^* . Consequently, a measure of inaccuracy that is aimed to assess the value of a given prognostic classification scheme should compare the estimated event-free probabilities with the observed ones.

Assume for now there is no censoring. The aim is to compare estimated event-free probabilities $\hat{S}(t^*|Z = z)$ for patients with covariates $Z = z$ to observed indicators of survival $I(T > t^*)$ leading to

$$BS(t^*) = \frac{1}{n} \sum_{i=1}^n \left(I(T_i > t^*) - \hat{S}(t^*|Z = z_i) \right)^2$$

where the sum is over all n patients. Models are preferred that minimize this quantity, also known as the quadratic score. Multiplication by a factor of two yields the Brier score, which was originally developed for judging the inaccuracy of probabilistic weather forecasts.^{131–133} The expected value of the Brier score may be interpreted as a mean square error of prediction if the event status at t^* is predicted by the estimated event-free probabilities. The Brier score takes values between zero and one; a trivial, constant prediction $\hat{S}(t^*) = 0.5$ for all patients yields a Brier score equal to 0.25.

If some closer relationship to the likelihood is intended, the so-called logarithmic score may be preferred, given by

$$LS(t^*) = -\frac{1}{n} \sum_{i=1}^n \left\{ I(T_i > t^*) \log \hat{S}(t^* | Z = z_i) + I(T_i \leq t^*) \log (1 - \hat{S}(t^* | Z = z_i)) \right\},$$

where we adopt conventions “ $0 \cdot \log 0 = 0$ ” and “ $1 \cdot \log 0 = -\infty$ ” (133,134). Again, models are preferred that minimize this score.

If we do not wish to restrict ourselves to one fixed time point t^* , we can consider the Brier and logarithmic scores as a function of time for $0 \leq t \leq t^*$. In case of the Brier score, we use the term *prediction error curve* for this function. If one wants a single number summary, this function can also be averaged over time by integrating it with respect to some weight function $W(t)$ over $t \in [0, t^*]$.¹²⁴

Censoring can be accommodated by reweighting the individual contributions in a similar way as in the calculation of the Kaplan–Meier estimator, so that consistency of estimates is preserved. The reweighting of uncensored observations and of observations censored after t is done by the reciprocal of the Kaplan–Meier estimate of the censoring distribution, whereas observations censored before t get weight zero. With this weighting scheme, a Brier or a logarithmic score under random censorship can be defined that enjoys the desirable statistical properties.¹²⁴ R^2 -type measures^{135–138} can also be readily defined by relating the Brier or logarithmic scores to the pooled Kaplan–Meier estimate, which is used as a universal prediction for all patients. Recently, a R^2 -type measure based on the improved measure of separation¹²² has also been developed.

We calculated the Brier score for the data of the GBSG-2 study. In [Figure 17.11\(a\)](#), the estimated prediction error curves of the classification schemes considered in Section 17.8 ([Table 17.8\(a\)](#)) are contrasted to the estimated prediction error curve based on the pooled Kaplan–Meier estimator. Because the latter prediction ignores all covariate information, it yields a benchmark value. It can be seen that the simplified COX index performs better than the NPI, and some further improvement is achieved by the CART index. Relative to the prediction with the pooled Kaplan–Meier estimate for all patients, there is only a moderate gain of accuracy.

In general, it has to be acknowledged that measures of inaccuracy tend to be large reflecting that predictions are far from being perfect.¹³⁹ In addition, it has to be mentioned that there will be overoptimism when a measure of inaccuracy is calculated from the same data where the prognostic classification scheme was derived, such as the case for the curves in [Figure 17.11\(a\)](#). To demonstrate this, [Figure 17.11\(b\)](#) shows prediction error curves using the external Freiburg DNA study for the same prognostic factors. As already indicated in [Table 17.8\(a\)](#), there is almost no discrimination between the various classification schemes in these independent test data at least for

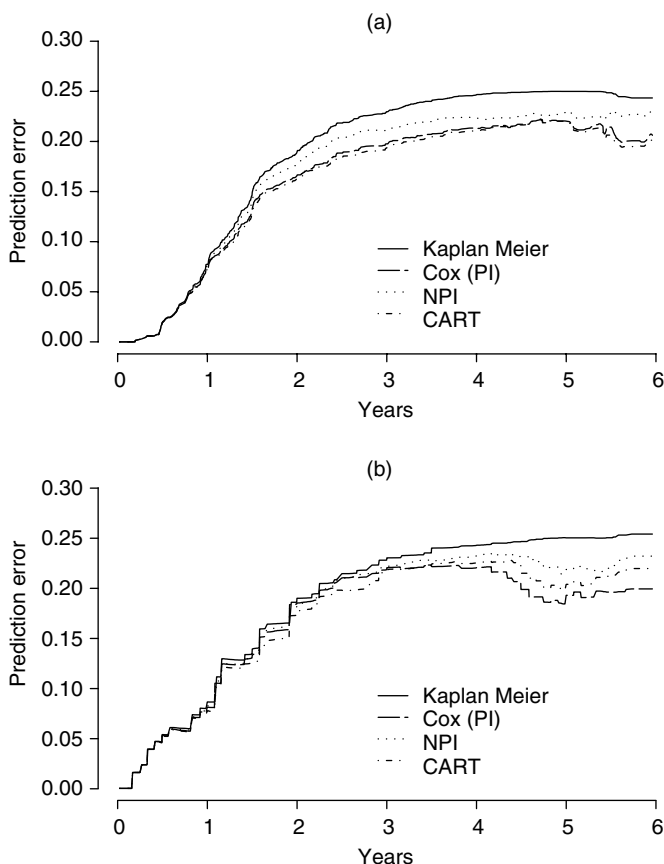


FIGURE 17.11 Estimated prediction error curves for various prognostic classification schemes derived in the data of the GBSG-2 study (a) and validated in the Freiburg DNA study (b).

values of t up to about 3 years. Afterward the classification scheme based on the COX model has slightly lower prediction error than the others.

In order to expand that issue further, we estimated prediction error curves for the various neural networks discussed in Section 17.8. The prediction error curves corresponding to the neural networks presented in Table 17.8(b) are displayed in Figure 17.12. The neural network with 20 hidden nodes and no weight decay has the lowest prediction error (Figure 17.12(a)), whereas the other three neural networks have a prediction error of similar magnitude to the other prognostic classification schemes. The resulting overoptimism can best be illustrated by inspection of Figure 17.12(b) where the Freiburg DNA study is used for validation. Here, the prediction error curve of the neural network with 20 hidden nodes and no weight decay used is highest, even exceeding the pooled Kaplan–Meier estimate. The two neural networks where weight decay is used have similar prediction error as the other prognostic classification schemes displayed in Figure 17.11(b).

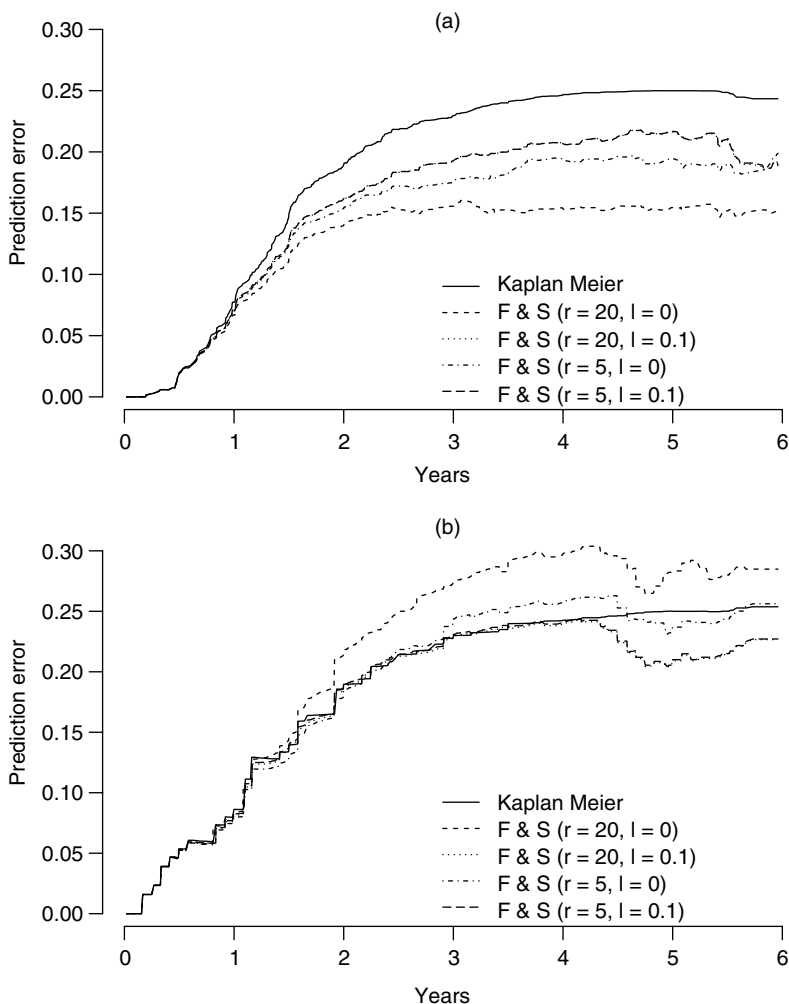


FIGURE 17.12 Estimated prediction error curves for various Faraggi-Simon neural networks derived in the data of the GBSG-2 study (a) and validated in the Freiburg DNA study (b).

To reduce the inherent overoptimism, cross-validation and resampling techniques may be employed in a similar way as for the estimation of error rates^{140–142} or for the reduction of bias of effect estimates as outlined in Section 17.4. For definitive conclusions, however, the determination of measures of inaccuracy in an independent test data set is absolutely necessary.⁹⁶

17.10 SAMPLE SIZE CONSIDERATIONS

To investigate the role of a new prognostic factor, careful planning of an appropriate study is required. Sample size and power formulae in survival analysis have been

developed for randomized treatment comparisons, but in the analysis of prognostic factors, the covariates included are expected to be correlated with the factor of primary interest. In this situation, the existing sample size and power formulae are not valid and may not be applied. In this section we give an extension of Schoenfeld's formula¹⁴³ to the situation where a correlated factor is included in the analysis.

Suppose we wish to study the prognostic relevance of a certain factor, denoted by Z_1 , in the presence of a second factor Z_2 , where either can be a composite score based on several factors. The criterion of interest is survival or event-free survival of the patients. We assume that the analysis of the main effects of Z_1 and Z_2 is performed with the Cox proportional hazards model given by

$$\lambda(t | Z_1, Z_2) = \lambda_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2)$$

where $\lambda_0(t)$ denotes an unspecified baseline hazard function and β_1 and β_2 are the unknown regression coefficients representing the effects of Z_1 and Z_2 . For simplicity we assume that Z_1 and Z_2 are binary with $p = P(Z_1 = 1)$ denoting the prevalence of $Z_1 = 1$. Assume that the effect of Z_1 shall be tested by an appropriate two-sided test based on the partial likelihood derived from the Cox model with significance level α and power $1 - \beta$ to detect an effect that is given by a hazard ratio of $\theta_1 = \exp(\beta_1)$.

For independent Z_1 and Z_2 , it was shown by Schoenfeld¹⁴³ that the total number of patients required is given by

$$N = \frac{(u_{1-\alpha/2} + u_{1-\beta})^2}{(\log \theta_1)^2 \psi(1-p)p},$$

where ψ is the probability of an uncensored observation and u_γ denotes the γ -quantile of the standard normal distribution. This is the same formula as that used by George and Desu,¹⁴⁴ Bernstein and Lagakos,¹⁴⁵ and Schoenfeld¹⁴⁶ in related problems.

The sample size formula depends on p , the prevalence of $Z_1 = 1$. The expected number of events, often also called the effective sample size, to achieve a prespecified power is minimal for $p = 0.5$, the situation of a randomized clinical trial with equal probabilities for treatment allocation. By using the same approximations as Schoenfeld,¹⁴³ one can derive a formula for the case when Z_1 and Z_2 are correlated with correlation coefficient ρ ; for details we refer to Schmoor et al.¹⁴⁷ This formula reads

$$N = \frac{(u_{1-\alpha/2} + u_{1-\beta})^2}{(\log \theta_1)^2 \psi(1-p)p} \left(\frac{1}{1 - \rho^2} \right)$$

where the factor $1/(1 - \rho^2)$ is the variance inflation factor (VIF).

This formula is identical to a formula derived by Lui¹⁴⁸ for the exponential regression model in the case of no censoring and to that developed by Palta and Amini¹⁴⁹ for the situation that the effect of Z_1 is analyzed by a stratified logrank test where $Z_2 = 0$ and $Z_2 = 1$ define the two strata. Table 17.9 gives, for some situations, the value of the VIF and the effective sample size $N\psi$, i.e., the number of events required to obtain a power of 0.8 to detect an effect to Z_1 of magnitude θ_1 . It shows

that the required number of events for the case of two correlated factors may increase up to a factor of 50% in situations realistic in practice.

The sample size formulae given above will now be illustrated by means of the GBSG-2 study. Suppose we want to investigate the influence of the progesterone receptor in the presence of tumor grade. The Spearman correlation coefficient of these two factors is $\rho = -0.377$; if they are categorized as binary variables we obtain $\rho = -0.248$ (Table 17.10). Taking the prevalence of progesterone-positive tumors, $p = 60\%$, into account, 213 events are required to detect a hazard ratio of 0.67, and 74 events are required to detect a hazard ratio of 0.5 with power 80% and significance level $\alpha = 5\%$. In this situation, the variance inflation factor is equal to 1.07 indicating that the correlation between the two factors has only little influence on power and required sample sizes.

If we want to investigate the prognostic relevance of progesterone receptor in the presence of estrogen receptor, a higher correlation must be considered. The Spearman correlation coefficient is $\rho = 0.598$ if both factors are measured on a quantitative scale and $\rho = 0.536$ if they are categorized into positive (>20 fmol) and negative (≤ 20 fmol) as given in Table 17.10. This leads to a variance inflation factor of 1.41 and a number of events of 284 and 97 required to detect a hazard ratio of 0.67 and

TABLE 17.9
Variance Inflation Factors and Effective Sample Size Required to Detect an Effect of Z_1 of Magnitude θ_1 with Power 0.8 as Calculated by the Approximate Sample Size Formula for Various Values of p, ρ, θ_1 ($\alpha = 0.05$)

p	ρ	VIF	$\theta_1 = 1.5$	$\theta_1 = 2$	$\theta_1 = 4$
			N_{Ψ}		
0.5	0	1	191	65	16
	0.2	1.04	199	68	17
	0.4	1.19	227	78	19
	0.6	1.56	298	102	26
0.3	0	1	227	78	19
	0.2	1.04	237	81	20
	0.4	1.19	271	93	23
	0.6	1.56	355	122	30

TABLE 17.10
Distribution of Progesterone Receptor by Tumor Grade and Estrogen Receptor in the GBSG-2 Study

		Tumor Grade		Estrogen Receptor	
		1	2 + 3	<20	≥ 20
Progesterone receptor	<20	5	264	190	79
	≥ 20	76	341	72	345
Correlation coefficient		-0.248		0.536	

0.5, respectively (power = 80%, significance level $\alpha = 5\%$). This has to be contrasted with the situation that both factors under consideration are uncorrelated; in this case the required number of events is 201 to detect a hazard ratio of 0.67 and 69 to detect a hazard ratio of 0.5, both with a power of 80% at a significance level of 5%.

From this aspect, the GBSG-2 study with 299 events does not seem too small to investigate the relevance of prognostic factors that exhibit at least a moderate effect (hazard ratio of 0.67 or 1.5). The question is whether it is large enough to permit the investigation of several prognostic factors. There have been some recommendations in the literature, based on practical experience or on results from simulation studies, regarding the event per variable relationship.^{68,150–153} More precisely, it is the number of events per model parameter that matters but is often overlooked. These recommendations range from 10 to 25 events per model parameter to ensure stability of the selected model and of corresponding parameter estimates and to avoid serious overfitting.

The sample size formula given above addresses the situation of two binary factors. For more general situations, such as factors occurring on several levels or factors with continuous distributions, the required sample size may be calculated using a more general formula that can be developed along the lines of Lubin and Gail.¹⁵⁴ The anticipated situation then has to be specified in terms of the joint distribution of the factors under study and the size of corresponding effects on survival. It may be more difficult to pose the necessary assumptions than in the situation of only two binary factors. Numerical integration techniques are then required to perform the necessary calculation.

If there are several factors in the analysis, one practical solution is to prespecify a prognostic score based on the existing standard factors as the second covariate to be adjusted for. Another possibility would be to adjust for the prognostic factor with the largest effect on survival and for which highest correlation is anticipated. Finally, it should be mentioned that a sample size formula for the investigation of interactive effects of two prognostic factors is also available.^{155,156,147}

17.11 CONCLUDING REMARKS

In this chapter we have considered statistical aspects of the evaluation of prognostic factors. Some general conclusions can be summarized as follows. A multivariate approach is absolutely essential. Thoughtful application of model building techniques should help to obtain models that are as simple and parsimonious as possible and to avoid serious overfitting in order to achieve generalizability for future patients. Thus, validation in an independent study is a further essential step. Some insight into the stability and generalizability of the derived models can be gained by cross-validation and resampling methods that, however, cannot be regarded to completely replace an independent validation study. For a concrete study, the statistical analysis should be carefully planned step by step and the model building process should at least principally be fixed in advance in a statistical analysis plan as is required in much more detail for clinical trials according to international guidelines.^{157,158}

There are a number of important topics that have not been mentioned or have been mentioned only in passing in this chapter. One of these topics is concerned with the handling of missing values in prognostic factors.^{159–163} We have also assumed

that effects of prognostic factors are constant over time and that prognostic factors are recorded and known at time of diagnosis. These assumptions do not cover the situation of time-varying effects and of time-dependent covariates. If multiple end-points or different events are of interest, the use of competing risk and multistate models may be indicated. For these topics that are also of importance for prognostic factor studies, we refer to more advanced textbooks in survival analysis^{15,24,26, 28} and current research papers. In general, the methods and approaches presented in this contribution have at least in part been selected and assessed according to the subjective views of the authors. Thus, other approaches might be seen as useful and adequate, too. What should not be a matter of controversy, however, is the need for a careful planning, conducting, and analyzing of prognostic factor studies in order to arrive at generalizable and reproducible results that could contribute to a better understanding and possibly to an improvement of the prognosis of cancer patients.

ACKNOWLEDGMENT

We would like to thank our colleagues Thomas Gerds, Erika Graf, and Claudia Schmoor for valuable contributions and Regina Gsellinger for her assistance in preparing the manuscript.

REFERENCES

1. McGuire WL. Breast cancer prognostic factors: evaluation guidelines. *J Natl Cancer Inst* 1991;83:154–155.
2. Infante-Rivard C, Villeneuve J-P, Esnaola S. A framework for evaluating and conducting prognostic studies: an application to cirrhosis of the liver. *J Clin Epidemiol* 1989;42:791–805.
3. Clark GM, ed. Prognostic factor integration. Special issue. *Breast Cancer Res Treat* 1992;22:185–293.
4. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *Br J Cancer* 1994;69:979–985.
5. Altman DG, Lyman GH. Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Res Treat* 1998;52:289–303.
6. Wyatt JC, Altman DG. Prognostic models: clinically useful or quickly forgotten? Commentary. *BMJ* 1995;311:1539–1541.
7. Armitage P, Gehan EA. Statistical methods for the identification and use of prognostic factors. *Int J Cancer* 1974;13:16–36.
8. Byar DP. Analysis of survival data: Cox and Weibull models with covariates. In: Mike V, Stanley KE, eds. *Statistics in Medical Research*. New York: Wiley, 1982:365–401.
9. Byar DP. Identification of prognostic factors. In: Buyse ME, Staquet MJ, Sylvester RJ, eds. *Cancer Clinical Trials: Methods and Practice*. Oxford: Oxford University Press, 1984:423–443.
10. Simon R. Use of regression models: statistical aspects. In: Buyse ME, Staquet MJ, Sylvester RJ, eds. *Cancer Clinical Trials: Methods and Practice*. Oxford: Oxford University Press, 1984:444–466.
11. George SL. Identification and assessment of prognostic factors. *Semin Oncol* 1988; 5:462–471.

12. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using “optimal” cut-points in the evaluation of prognostic factors. Commentary. *J Natl Cancer Inst* 1994; 86:829–835.
13. Altman DG, De Stavola BL, Love SB, Stepniowska KA. Review of survival analyses published in cancer journals. *Br J Cancer* 1995;72:511–518.
14. Gospodarowicz MK, Henson DE, Hutter RVP, O’Sullivan B, Sobin LH, Wittekind C. *Prognostic Factors in Cancer*. 2nd ed. Heidelberg, New York: Springer, 2001.
15. Marubini E, Valsecchi MG. *Analysing Survival Data from Clinical Trials and Observational Studies*. Chichester: Wiley, 1995.
16. Parmar MKB, Machin D. *Survival Analysis: A Practical Approach*. Chichester: Wiley, 1995.
17. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gotzsche PC, Lang T for the CONSORT Group. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;134:663–694.
18. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Reporting recommendations for tumor marker prognostic studies. 2004 (submitted).
19. Riley RD, Abrams KR, Sutton AJ, Lambert PC et al. Reporting of prognostic markers: current problems and development of guidelines for evidence based practice in the future. *Br J Cancer* 2003;88:1191–1198.
20. Altman DG. Systematic reviews of studies of prognostic variables. In: Egger M, Davey S, Altman D eds. *Systematic Reviews in Health Care: Meta Analysis in Context*. London: BHY Publ. group 2001:228–247.
21. Harris EK, Albert A. *Survivorship Analysis for Clinical Studies*. New York: Marcel Dekker, 1991.
22. Lee ET. *Statistical Methods for Survival Data Analysis*. 2nd ed. New York: Wiley, 1992.
23. Collett D. *Modelling Survival Data in Medical Research*. London: Chapman & Hall, 1994.
24. Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer, 1997.
25. Therneau TM, Grambsch PM. *Modeling Survival Data*. New York: Springer, 2000.
26. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. New York: Wiley, 1980.
27. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*. New York: Wiley, 1991.
28. Andersen PK, Borgan O, Gill RD, Keiding N. *Statistical Methods Based on Counting Processes*. New York: Springer, 1992.
29. Burke HB, Henson DE. Criteria for prognostic factors and for an enhanced prognostic system. *Cancer* 1993;72:3131–3135.
30. Henson DE. Future directions for the American Joint Committee on Cancer. *Cancer* 1992;69:1639–1644.
31. Fielding LP, Fenoglio-Preiser CM, Freedman LS. The future of prognostic factors in outcome prediction for patients with cancer. *Cancer* 1992;70:2367–2377.
32. Fielding LP, Henson DE. Multiple prognostic factors and outcome analysis in patients with cancer. *Cancer* 1993;71:2426–2429.
33. Rothman KJ, Greenland S. *Modern Epidemiology*. 2nd ed. Philadelphia: Lippincott-Raven Publishers, 1998.
34. Simon R. Patients subsets and variation in therapeutic efficacy. *Br J Cancer* 1982; 14:473–482.

35. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 1985;41:361–372.
36. Byar DP. Assessing apparent treatment covariate interactions in randomized clinical trials. *Stat Med* 1985;4:255–263.
37. Schmoor C, Ulm K, Schumacher M. Comparison of the Cox model and the regression tree procedure in analyzing a randomized clinical trial. *Stat Med* 1993; 12:2351–2366.
38. Royston P, Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Stat Med* 2004;23:2509–2525.
39. Bloom HJG, Richardson WW. Histological grading and prognosis in primary breast cancer. *Br J Cancer* 1957;2:359–377.
40. Pfisterer J, Kommoss F, Sauerbrei W, Menzel D, Kiechle M, Giese E, Hilgarth M, Pfliederer A. DNA flow cytometry in node positive breast cancer: prognostic value and correlation to morphological and clinical factors. *Anal Quant Cytol Histol* 1995; 17:406–412.
41. Schumacher M, Bastert G, Bojar H, Hübner K, Olschewski M, Sauerbrei W, Schmoor C, Beyerle C, Neumann RLA, Rauschecker HF for the German Breast Cancer Study Group. Randomized 2×2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *J Clin Oncol* 1994; 12:2086–2093.
42. Schmoor C, Olschewski M, Schumacher M. Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies. *Stat Med* 1996;15:263–271.
43. Cox DR. Regression models and life tables (with discussion). *J R Stat Soc, Ser. B* 1972;34:187–220.
44. Schumacher M, Holländer N, Sauerbrei W. Reduction of bias caused by model building. *Proceedings of the Statistical Computing Section, American Statistical Association*, 1996; 1–7.
45. Lausen B, Schumacher M. Maximally selected rank statistics. *Biometrics* 1992; 48:73–85.
46. Hilsenbeck SG, Clark GM. Practical P-value adjustment for optimally selected cut-points. *Stat Med* 1996; 15:103–112.
47. Worsley KJ. An improved Bonferroni inequality and applications. *Biometrika* 1982; 69:297–302.
48. Lausen B, Sauerbrei W, Schumacher M. Classification and regression trees (CART) used for the exploration of prognostic factors measured on different scales. In: Dirschedl P, Ostermann R, eds. *Computational Statistics*. Heidelberg: Physica-Verlag, 1994:1483–1496.
49. Lausen B, Schumacher M. Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. *Comput Stat Data Analysis* 1996; 21:307–326.
50. Hothorn T, Lausen B. On the exact distribution of maximally selected rank statistics. *Comput Stat Data Analysis* 2003; 121–137.
51. Verweij P, Van Houwelingen HC. Cross-validation in survival analysis. *Stat Med* 1993;12:2305–2314.
52. Schumacher M, Holländer N, Sauerbrei W. Resampling and cross-validation techniques: a tool to reduce bias caused by model building. *Stat Med* 1997; 16:2813–2827.
53. Van Houwelingen HC, Le Cessie S. Predictive value of statistical models. *Stat Med* 1990;9:1303–1325.

54. Holländer N, Sauerbrei W, Schumacher M. Confidence intervals for the effect of a prognostic factor after selection of an “optimal” cutpoint. *Stat Med* 2004; 23:1701–1713.
55. Andersen PK. Survival analysis 1982–1991: the second decade of the proportional hazards regression model. *Stat Med* 1991;10:1931–1941.
56. Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. *Appl Stat* 1999;48:313–329.
57. Miller AJ. *Subset Selection in Regression*. London: Chapman and Hall, 1990.
58. Teräsvirta T, Mellin I. Model selection criteria and model selection tests in regression models. *Scand J Stat* 1986;13:159–171.
59. Mantel N. Why stepdown procedures in variable selection? *Technometrics* 1970;12:621–625.
60. Sauerbrei W. Comparison of variable selection procedures in regression models: a simulation study and practical examples. In: Michaelis J, Hommel G, Wellek S, eds. *Europäische Perspektiven der Medizinischen Informatik, Biometrie und Epidemiologie*. MMV: München, 1993:108–113.
61. Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978;6:461–464.
62. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Appl Stat* 1994;43:429–467.
63. Holländer N, Schumacher M. Estimating the functional form of a continuous covariate’s effect on survival time. *Comput Stat Data Anal* 2005; in press.
64. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *J R Stat Soc, Ser. A*, 1999;162:71–94.
65. Sauerbrei W, Royston P, Bojar H, Schmoor C, Schumacher M and the German Breast Cancer Study Group (GBSG). Modelling the effects of standard prognostic factors in node positive breast cancer. *Br J Cancer* 1999; 79:1752–1760.
66. Harrell FE Jr. *Regression Modelling Strategies, with Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer, 2001.
67. Valsecchi MG, Silvestri D. Evaluation of long-term survival: use of diagnostics and robust estimators with Cox’s proportional hazards model. *Stat Med* 1996; 15:2763–2780.
68. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; 15:361–387.
69. Hastie TJ, Tibshirani RJ. *Generalized Additive Models*. New York: Chapman and Hall, 1990.
70. Chen CH, George SL. The bootstrap and identification of prognostic factors via Cox’s proportional hazards regression model. *Stat Med* 1985;4:39–46.
71. Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. *Stat Med* 1989;8:771–783.
72. Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. *Stat Med* 1992; 11:2093–2109.
73. Royston P, Sauerbrei W. Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. *Stat Med*, 2003; 22:639–659.
74. Breiman L, Friedman JH, Olshen R, Stone CJ. *Classification and Regression Trees*. Monterey: Wadsworth, 1984.
75. Zhang H, Crowley J, Sox HC, Olshen R. Tree-structured statistical methods. In: Armitage P, Colton T, eds. *Encyclopedia of Biostatistics*. Chichester: Wiley, 1998:4561–4573.

76. Gordon L, Olshen R. Tree-structured survival analysis. *Can Treatm Rep* 1985;69:1065–1069.
77. LeBlanc M, Crowley J. Relative risk regression trees for censored survival data. *Biometrics* 1992;48:411–425.
78. LeBlanc M, Crowley J. Survival trees by goodness of split. *JASA* 1993; 88:457–467.
79. Segal MR. Regression trees for censored data. *Biometrics* 1988; 44:35–47.
80. Segal MR. Tree-structured survival analysis in medical research. In: Everitt BS, Dunn G, eds. *Statistical Analysis of Medical Data: New Developments*. London: Arnold, 1998:101–125.
81. Ciampi A, Hendricks L, Lou Z. Tree-growing for the multivariate model: the RECPAM approach. In: Dodge Y, Whittaker J, eds. *Computational Statistics*. Vol 1. Heidelberg: Physica-Verlag, 1992: 131 – 136.
82. Tibshirani R, LeBlanc M. A strategy for binary description and classification. *J Comput Graph Stat* 1992;1:3–20.
83. Sauerbrei W. On the development and validation of classification schemes in survival data. In: Klar R, Opitz O, eds. *Classification and Knowledge Organization*. Berlin, Heidelberg, New York: Springer, 1997:509–518.
84. Breiman L. Bagging predictors. *Machine Learn* 1996;24:123–140.
85. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. *Ann Stat* 2000;2:334–374.
86. Breiman L. Random forests. *Machine Learn* 2001;45:5–32.
87. Diettrich TG. An experimental comparison of three methods for construting ensembles of decision trees: bagging, boosting, and randomization. *Machine Learn* 2000; 40:139–157.
88. Hothorn T, Lausen B, Benner A, Radespiel-Tröger M. Bagging survival trees. *Stat Med* 2003;
89. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat Sci* 2001;16:199–231.
90. Haybittle JL, Blamey RW, Elston CW, Johnson J, Doyle PJ, Campbell FC, Nicholson RI, Griffiths K. A prognostic index in primary breast cancer. *Br J Cancer* 1982; 45:361–366.
91. Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Res Treat* 1992;22:207–219.
92. Balslev I, Axelsson CK, Zedeler K, Rasmussen BB, Carstensen B, Mouridsen HT. The Nottingham Prognostic Index applied to 9,149 patients from the studies of the Danish Breast Cancer Cooperative Group (DBCG). *Breast Cancer Res Treat* 1994; 32:281–290.
93. Collett K, Skjaerven R, Machle BO. The prognostic contribution of estrogen and progesterone receptor status to a modified version of the Nottingham Prognostic Index. *Breast Cancer Res Treat* 1998;48:1–9.
94. Copas JB. Using regression models for prediction: shrinkage and regression to the mean. *Stat Methods Med Res* 1997;6:167–183.
95. Vach W. On the relation between the shrinkage effect and a shrinkage method. *Comput Stat* 1997;12:279–292.
96. Ripley BD. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press, 1996.
97. Baxt WG. Application of artificial neural networks to clinical medicine. *Lancet* 1995; 346:1135–1138.
98. Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. *Lancet* 1995; 346:1075–1079.
99. Dybowski R, Gant, V. Artificial neural networks in pathology and medical laboratories. *Lancet* 1995;346:1203–1207.

100. Wyatt J. Nervous about artificial neural networks? *Lancet* 1995;346:1175–1177.
101. Schwarzer G, Vach W, Schumacher M. On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Stat Med* 2000;19:541–561.
102. Penny W, Frost D. Neural networks in clinical medicine. *Med Decis Making* 1996; 16:386–398.
103. Cheng B, Titterton DM. Neural networks: a review from a statistical perspective (with discussion). *Stat Sci* 1994;9:2–54.
104. Ripley BD. Statistical aspects of neural networks. In: Barndorff-Nielsen OE, Jensen JL, eds. *Networks and Chaos: Statistical and Probabilistic Aspects*. London: Chapman and Hall, 1993: 40–123.
105. Schumacher M, Rossner R, Vach W. Neural networks and logistic regression: Part I. *Comput Stat Data Anal* 1996;21:661–682.
106. Stern HS. Neural networks in applied statistics (with discussion). *Technometrics* 1996;38:205–220.
107. Warner B, Misra M. Understanding neural networks as statistical tools. *American Statistician* 1996;50:284–293.
108. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. *Neural Networks* 1989;2:359–366.
109. Ripley RM. Neural network models for breast cancer prognosis. Ph.D. dissertation, University of Oxford, Dept. of Engineering Science, Oxford, 1998.
110. Liestøl K, Andersen PK, Andersen U. Survival analysis and neural nets. *Stat Med* 1994;13:1189–1200.
111. Ravdin PM, Clark GM. A practical application of neural network analysis for predicting outcome on individual breast cancer patients. *Breast Cancer Res Treat* 1992; 22:285–293.
112. Ravdin PM, Clark GM, Hilsenbeck SG, Owens MA, Vendely P, Pandian MR, McGuire WL. A demonstration that breast cancer recurrence can be predicted by neural network analysis. *Breast Cancer Res Treat* 1992;21:47–53.
113. De Laurentiis M, Ravdin PM. Survival analysis of censored data: neural network analysis detection of complex interactions between variables. *Breast Cancer Res Treat* 1994;32:113–118.
114. Burke HB. Artificial neural networks for cancer research: outcome prediction. *Semin Surg Oncol* 1994;10:73–79.
115. McGuire WL, Tandon AK, Allred DC, Chamness GC, Ravdin PM, Clark GM. Treatment decisions in axillary node-negative breast cancer patients. *Monogr Natl Cancer Inst* 1992;11:173–180.
116. Kappen HJ, Neijt JP. Neural network analysis to predict treatment outcome. *Ann Oncol* 1993;4(suppl):31–34.
117. Faraggi D, Simon R. A neural network model for survival data. *Stat Med* 1995; 14:73–82.
118. Biganzoli E, Boracchi P, Mariani L, Marubini E. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Stat Med* 1998;17:1169–1186.
119. Ripley BD, Ripley RM. Neural networks as statistical methods in survival analysis. In: Dybowski R, Gant V, eds. *Clinical Application of Artificial Neural Networks*. New York: Cambridge University Press, 2001:237–255.
120. Biganzoli E, Boracchi P, Marubini E. A general framework for neural network models on censored survival data. *Neural Networks* 2002;15:209–218.
121. Ripley RM, Harris AL, Tarassenko L. Non-linear survival analysis using neural networks. *Stat Med* 2004;23:825–842.
122. Sauerbrei W, Hübner K, Schmoor C, Schumacher M for the German Breast Cancer Study Group. Validation of existing and development of new prognostic classification

- schemes in node negative breast cancer. *Breast Cancer Res Treat* 1997; 42:149–163. Correction. *Breast Cancer Res Treat* 1998;48:191–192.
123. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004;23:723–748.
124. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* 1999;18:2529–2545.
125. Schumacher M, Graf E, Gerdts T. How to assess prognostic models for survival data: a case study in oncology. *Methods Inf Med* 2003;42:564–571.
126. Parkes MC. Accuracy of predictions of survival in later stages of cancer. *BMJ* 1972; 264:29–31.
127. Forster LA, Lynn J. Predicting life span for applicants to inpatient hospice. *Arch Intern Med* 1988;148:2540–2543.
128. Maltoni M, Pirovano M, Scarpi, E, Marinari, M, Indelli M, Arnoldi E, Galluci M, Frontini L, Piva L, Amadori D. Prediction of survival of patients terminally ill with cancer. *Cancer* 1995;75:2613–2622.
129. Henderson R, Jones M. Prediction in survival analysis. model or medic. In: Jewell NP, Kimber AC, Ting Lee ML, Withmore GA, eds. *Lifetime Data: Models in Reliability and Survival Analysis*. Dordrecht: Kluwer Academic Publishers, 1996: 125–129.
130. Henderson R. Problems and prediction in survival-data analysis. *Stat Med* 1995; 3:143–152.
131. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly Weather Rev* 1950;78:1–3.
132. Hilden J, Habbema JDF, Bjerregard B. The measurement of performance in probabilistic diagnosis III: methods based on continuous functions of the diagnostic probabilities. *Methods of Informations in Medicine* 1978;17:238–246.
133. Hand DJ. *Construction and Assessment of Classification Rules*. Chichester: Wiley, 1997.
134. Shapiro AR. The evaluation of clinical predictions. *N Engl J Med* 1997; 296:1509–1514.
135. Korn EJ, Simon R. Explained residual variation, explained risk and goodness of fit. *American Statistician* 1991;45:201–206.
136. Schemper M. The explained variation in proportional hazards regression. *Biometrika* 1990; 77:216–218. Correction. *Biometrika* 1994;81:631.
137. Graf E, Schumacher M. An investigation on measures of explained variation in survival analysis. *Statistician* 1995;44:497–507.
138. Schemper M, Stare J. Explained variation in survival analysis. *Stat Med* 1996; 15:1999–2012.
139. Ash A, Schwartz M. R2: a useful measure of model performance when predicting a dichotomous outcome. *Stat Med* 1999;18:375–384.
140. Efron B. Estimating the error rate of prediction rule: improvement on cross-validation. *JASA* 1983;78:316–330.
141. Efron B, Tibshirani R. Improvement on cross-validation: the .632+ bootstrap method. *JASA* 1997;92:548–560.
142. Wehberg S, Schumacher M. A comparison of nonparametric error rate estimation methods in classification problems. *Biometrical J.* 2004;46:35–47.
143. Schoenfeld DA. Sample size formula for the proportional-hazard regression model. *Biometrics* 1983;39:499–503.
144. 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.
145. Bernstein D, Lagakos SW. Sample size and power determination for stratified clinical trials. *J Stat Compu Sim* 1978;8:65–73.

146. Schoenfeld DA. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* 1981;68:316–319.
147. Schmoor, C, Sauerbrei, W, Schumacher M. Sample size considerations for the evaluation of prognostic factors in survival analysis. *Stat Med* 2000;19:441–452.
148. 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.
149. Paltu M, Amini SB. Consideration of covariates and stratification in sample size determination for survival time studies. *J Chronic Dis* 1985;38:801–809.
150. Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis: I. Background, goals and general strategy. *J Clin Epidemiol* 1995;48:1495–1501.
151. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards analysis: II. Accuracy and precision of regression estimates. *J Clin Epidemiol* 1995;48:1503–1510.
152. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modeling strategies for improved prognostic prediction. *Stat Med* 1984;3:143–152.
153. Harrell FE, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems and suggested solutions. *Cancer Treat Rep* 1985;69:1071–1077.
154. Lubin JH, Gail MJ. On power and sample size for studying features of the relative odds of disease. *Am J Epidemiol* 1990;131:551–566.
155. Peterson B, George SL. Sample size requirements and length of study for testing interaction in a $2 \times k$ factorial design when time-to-failure is the outcome. *Control Clin Trials* 1993;14:511–522.
156. Olschewski M, Schumacher M, Davis K. Analysis of randomized and nonrandomized patients in clinical trials using the comprehensive cohort follow-up study design. *Control Clin Trials* 1992;13:226–239.
157. European Community, CPMP Working Party on Efficacy of Medicinal Products. Biostatistical methodology in clinical trials in applications for marketing authorisations for medicinal products. Note for Guidance III/3630/92-EN, December 1994. *Stat Med* 1995;14:1659–1682.
158. International Conference of Harmonisation and Technical Requirements for Registration of Pharmaceuticals for Human Drugs. ICH Harmonised Tripartite Guideline: Guideline for Good Clinical Practice. Recommended for Adoption at Step 4 of the ICH Process on 1 May 1996 by the ICH Steering Committee.
159. Vach, W. *Logistic Regression with Missing Values in the Covariates*. Lecture Notes in Statistics 86. New York: Springer 1994.
160. Vach, W. Some issues in estimating the effect of prognostic factors from incomplete covariate data. *Stat Med* 1997;16:57–72.
161. Robins JM, Rotnitzky A, Zhao LD. Estimation of regression coefficients when some regressors are not always observed. *JASA* 1994;89:846–866.
162. Lipsitz SR, Ibrahim JG. Estimating equations with incomplete categorical covariates in the Cox model. *Biometrics* 1998;54:1002–1013.
163. Buuren van S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* 1999;18:681–694.

18 Statistical Methods to Identify Predictive Factors

*Kurt Ulm, Monika Kriner, Sonja Eberle,
Martin Reck, and Sybill Hessler*

CONTENTS

18.1	Introduction	335
18.2	Methods	337
18.2.1	Cox Model (Prognostic Model)	337
18.2.2	Fractional Polynomials	337
18.2.3	Identification of Predictive Factors (Bump Hunting)	338
18.2.3.1	Our Proposal of the Approach	338
18.3	Lung Cancer Example	339
18.3.1	Description of the Data	339
18.3.2	Construction of the Prognostic Model.....	340
18.3.3	Bump Hunting (Identification of Predictive Factors)	342
18.4	Final Remarks	343
	References	344

18.1 INTRODUCTION

In randomized clinical trials a new therapy is usually compared with the standard treatment or with placebo. Within the intention-to-treat analysis, all patients are used for the overall comparison. But independent of the overall result, the clinicians want to know whether there are subsets of patients with different outcomes. If the new therapy is superior to the standard, the question is whether there are patients who will not benefit from or may be harmed by the new treatment. The same holds if there is no advantage to the new therapy. There may well be patients who will benefit or may be harmed by the new drug. The background of these concerns is the tendency in recent years to select the appropriate treatment for a specific patient based on his or her prognostic or predictive factors.

In cancer, the information about the prognosis based on classical factors, such as the TNM-classification, may be limited and new factors should be identified.¹ An example for breast cancer is the research on the impact of the urokinase-type plasminogen activator and its inhibitor.²

In the following we will assume that in a clinical trial potential prognostic and predictive factors are available. Prognostic factors are covariates that have an influence on the prognosis of a patient independently from the treatment. We use the term *predictive* to describe an interaction between a factor and the treatment.³

The main question in this context is why a patient under the new therapy lives longer or shorter than the average or than patients under the standard treatment. Several reasons may be possible. First of all, it could be by chance, i.e., we have no explanation for the better performance under the new therapy. Secondly, the patient could have a better or worse prognosis than the average due to prognostic factors. Thus, a good prognostic model can find such factors and may be helpful in answering this question. Finally, the therapy could really be working, i.e., there are patients with a positive or negative reaction to the new treatment. In such cases predictive factors may be responsible for the difference in survival.

The identification of these factors has been addressed by numerous authors. The approach mostly used to identify predictive factors is to extend a regression model with interaction terms between the treatment and certain factors. This analysis is simple if a factor is categorical. For the analysis of continuous factors, the form of the relationship has to be taken into account. Royston and Sauerbrei³ suggest the use of fractional polynomials. Bonetti and Gelber⁴ explore the possible interaction by using their STEPP procedure in identifying subpopulations by either a *sliding window* or *tail-oriented* approach based on the comparison of both treatment groups with increasing level of the continuous factor. But both methods are restricted to first order interactions. In a regression model interactions of higher order can be taken into consideration. However, the problem is which combinations should be considered and how. Ciampi et al.⁵ presented an extension of CART,⁶ which was primarily designed for prognostic classification. This approach starts with the whole set of data. The idea is to divide the patients into two groups by maximizing the likelihood ratio statistics of two models, one model with treatment as the only factor, and the other with different treatment effects in both subgroups. All possible splits are considered. Each of the two subgroups is investigated for further splits. In the first step the regression model is as described above with an interaction term. The method by Ciampi et al. is one way to reduce the number of all possible interactions.

Another approach is described by O'Quigley and Natarajan.⁷ They use a change-point model for describing the time-varying effect of a prognostic factor. The authors refer to their method as a brutal approximation to reality. The same approach can be transferred to identify predictive factors. This model is equivalent to the approach presented by Ciampi et al.⁵ but restricted to first order interactions.

The goal of this paper is to propose a statistical procedure to identify predictive factors in the context of survival data. The procedure contains two steps. First, prognostic factors are identified on one part of the data using only the patients under the standard treatment or placebo. This analysis results in a prognostic model. In the second step this prognostic model is applied to the other half of the data, the patients under the new drug. Patients receiving the new therapy who are not well described by the prognostic model are investigated for predictive factors. To obtain a prognostic model we use the Cox model with some extensions such as fractional polynomials. To obtain a predictive model we use bump hunting. Data from a study on lung cancer will be used to illustrate these methods.

18.2 METHODS

18.2.1 COX MODEL (PROGNOSTIC MODEL)

The most widespread method used to analyse survival data is the Cox model.⁸ Denoting the time of follow-up by t and potential prognostic factors by $z = (z_j, \dots, z_p)$, the model is usually given in the following form:

$$\lambda(t | z) = \lambda(t | z = 0) \cdot \exp\left(\sum_{j=1}^p f_j(z_j)\right) = \lambda_0(t) \cdot \exp\left(\sum_{j=1}^p f_j(z_j)\right), \quad (18.1)$$

with $\lambda_0(t)$ being the baseline hazard rate and $\lambda(t | z)$ the hazard rate at time t given the factors z . The influence of a single factor z_j on the hazard can be modeled as

$$f_j(z_j) = \begin{cases} \beta_j z_j & \text{for a linear relationship or for categorical variables} \\ \beta_j(z_j) & \text{for non-linear effects of continuous factors} \\ \beta_j(t) z_j & \text{for time-varying effects (proportional hazards} \\ & \text{assumption not fulfilled)} \end{cases} \quad (18.2)$$

The hazard ratio (HR) or relative risk (RR) is defined as

$$\text{HR}(z) = \frac{\lambda(t | z)}{\lambda_0(t)} = \exp\left(\sum_{j=1}^p f_j(z_j)\right) \quad (18.3)$$

18.2.2 FRACTIONAL POLYNOMIALS

Suppose for the moment we have a single continuous factor z . Then the influence of this factor on the outcome can be modeled as a straight line ($= \beta \cdot z$) or as a function $\beta(z)$. Two main approaches described in the literature for how to select $\beta(z)$ are smoothing splines⁹ and its extensions¹⁰ and fractional polynomials.¹¹ In the latter, the idea is to construct a function $\beta(z)$ consisting of some polynomials of the form z^p , with the exponent p out of a given set of 8 values ($p \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ with $p = 0$ identical to $\ln z$, the natural logarithm of z). Either one term or at most two terms seem to be sufficient for most practical applications, i.e., $\beta(z) = \beta \cdot z^p$ or $\beta(z) = \beta_1 \cdot z^{p_1} + \beta_2 \cdot z^{p_2}$.^{3,11} Royston and Altman¹¹ ascertained that a model with $m > 2$ terms rarely yields any improvement compared to models with 1 or 2 terms. The best model is determined due to some goodness-of-fit criterion, i.e., the model with the smallest p -value of the likelihood ratio statistic is chosen, adjusted for the appropriate number of degrees of freedom ($DF = 2$ for $m = 1$ or $DF = 4$ for $m = 2$). The advantage of this approach is the representation of $\beta(z)$ in a functional form. Furthermore, standard software is available to compute fractional polynomials (e.g., STATA).

The same approach can be used in investigating the assumption of proportional hazards.¹² Considering again one factor z , the idea is to extend the linear assumption $\beta \cdot z$ into $\beta(t) \cdot z$. Now the influence of the factor z on the hazard ratio can be described as a function of time, which can be analyzed by standard software for time-dependent covariates by specifying $\beta(t) \cdot z = \beta \cdot t^p \cdot z = \beta \cdot z(t)$ with $z(t) = z \cdot t^p$. The

hypotheses of interest are whether $\beta(t)$ is constant ($\beta(t) = \beta$) or different from zero ($\beta(t) = 0$).

18.2.3 IDENTIFICATION OF PREDICTIVE FACTORS (BUMP HUNTING)

Predictive factors are covariates that interact with treatment in relation to the outcome. In a regression model, like the Cox model, an interaction term $x \cdot z$ can be added, with x denoting the treatment and z a possible predictive factor. The analysis is straightforward if z is a binary factor or has categorical outcomes. However, if z is continuous, then the form of the relationship has to be taken into account. Again fractional polynomials can be used,

$$f(x, z) = \beta \cdot x + f_1(z) + f_2(z)x \quad (18.4)$$

$f_1(z)$ and $f_2(z)$ being defined as in (18.2).

The hypothesis of interest is whether $f_2(z) = 0$, which indicates no interaction between x and z .

For more predictive factors, the analysis becomes more cumbersome if interactions of higher orders have to be investigated. In this situation an approach from data mining may be helpful, such as the bump hunting procedure introduced by Friedman and Fisher.¹³ Bump hunting is one method to detect subgroups of patients with different responses to the new treatment. Therefore, this method seems to be suitable to identify predictive factors.

In the following the whole procedure and especially bump hunting are described in greater detail.

18.2.3.1 Our Proposal of the Approach

The idea of the approach is to split the data and use half of the data, the patients under the standard treatment, to create a good prognostic model. This model is then applied to the other half of the data, the patients under the new treatment. In this group we look for similarities among those patients whose outcomes are not well described by the prognostic model. The deviation from the prognostic model is measured by martingale residuals M_j ,^{14–16}

$$M_j = \delta_j - \hat{\Lambda}_j(t_j, z_j), \quad -\infty \leq M_j \leq 1 \quad (18.5)$$

with δ_j the survival status (0 = censored, 1 = dead) and $\hat{\Lambda}_j(t_j, z_j)$ the estimated cumulative hazard rate based on the prognostic model.

A correct prognosis of the prognostic model applied to the patients under the new therapy who are not used in establishing the model would result in residuals around 0. Patients who live longer than predicted by the prognostic model have large negative martingale residuals. These patients may benefit from the new treatment:

$$\delta_j = 0 \text{ and } \hat{\Lambda}_j > 0 \text{ or } \delta_j = 1 \text{ and } \hat{\Lambda}_j \gg 1 \Rightarrow M_j < 0$$

On the other hand there will be patients who live shorter than expected. These patients will have positive martingale residuals and may be harmed by the new therapy:

$$\delta_j = 1 \text{ and } \hat{\Lambda}_j < 1 \Rightarrow M_j > 0$$

We are interested in patients with either high ($M > 0$) or low ($M < 0$) martingale residuals. Bump hunting can detect subgroups of patients with similar properties by stepwise maximizing or minimizing the mean of the martingale residuals.

Without loss of generality, we look for minimization ($M_j < 0$), i.e. for patients who may benefit from the new therapy. The algorithm successively restricts the data set by peeling away particular percentages of the data defined by simple logical rules (called *borders*) concerning the covariates (e.g., age < 35 years or gender = 1). In the remaining subset, the mean of the residuals is smaller than that of the whole data set. Through stepwise border selection the data set gets smaller, and the mean of the residuals decreases. At each step the remaining subset is compared to the corresponding subset in the group under standard treatment (with the same properties) by the log rank test.^{17,18} The process is repeated until the difference in survival is statistically significant or the number of the remaining patients gets too small. The intersection of all borders forms one *box*. This box is removed from the data set, and the procedure is repeated. If more than one box is found, the union of these boxes is called *bump*. The final bump represents the group of patients who may benefit from the new therapy.

The same procedure can be performed in maximizing the mean of the martingale residuals in order to identify patients who may be harmed by the new treatment.

To increase stability of the algorithm, Kehl¹⁹ introduced bootstrapping. A predefined number of bootstrap samples is drawn from the original data set, and bump hunting is applied to all these samples as well as to the original data. The border that is selected most frequently by the algorithm is finally chosen as the actual border. This selection is repeated in each step of the bump hunting procedure.

18.3 LUNG CANCER EXAMPLE

18.3.1 DESCRIPTION OF THE DATA

In this study 608 patients with small-cell lung cancer have been enrolled.²⁰ A new combination of chemotherapy composed of the substances Paclitaxel (Taxol), Etoposide, and Carboplatin (TEC, $n = 301$) is compared to the classical combination of the substances Carboplatin, Etoposide, and Vincristine (CEV, $n = 307$). The follow-up period is from January 1998 to December 1999. The median survival time is 11.7 months (12.2 months in the experimental arm and 11.3 months in the classical arm). 84.7% of the patients died (81.1% under TEC and 88.3% under CEV). [Figure 18.1](#) shows the Kaplan–Meier curves for the whole sample, separated for the two treatment arms. The log rank test ($p(LR) = 0.0237$) detects a significantly better performance of the new treatment TEC concerning the survival time ($HR = 1.22$; 95% CI: 1.03–1.45).

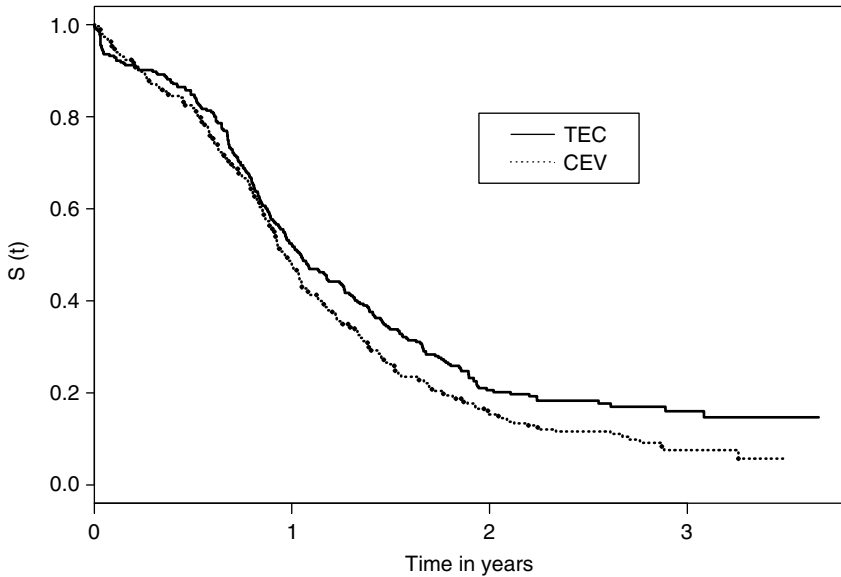


FIGURE 18.1 Survival curves of the lung cancer data.

18.3.2 CONSTRUCTION OF THE PROGNOSTIC MODEL

For the analysis 601 patients with full data have been used (TEC: $n = 296$, CEV: $n = 305$). Table 18.1 gives a short description of the potential prognostic and predictive factors used in the analysis. All categorical variables were dummy coded with the first category as reference. In order to derive the prognostic model, we selected all patients treated with the classical therapy (CEV). A statistically significant influence of gender ($p = 0.01$), lymph nodes ($p = 0.029$) and stage ($p < 0.001$) could be detected.

We observed a time-varying effect of the factor *stage*. Table 18.2 shows the result: the use of fractional polynomials yielded a function of time and the corresponding p -value for testing the time-varying effect of stage adjusted for gender and lymph node affection.

Figure 18.2 displays the effect of stage: in the first two years, patients with an extended disease have a higher mortality rate; after that, the patients with limited disease have the higher risk. The left panel shows the Kaplan–Meier curves for the variable *stage*, and the right panel is the resulting optimal fractional polynomial.

In order to improve the goodness of fit of the prognostic model, we introduced first order interactions of all factors. However, no statistically significant interaction could be identified. Hence, the linear predictor of the final prognostic model adds up to

$$\begin{aligned} \sum (f_i(z_j)) = & 0.651 \cdot (1.005 \cdot t - 0.213 \cdot t^2) \text{stage} + 0.396 \cdot \text{gender} \\ & + 1.332 \cdot N1 + 1.659 \cdot N2 + 1.678 \cdot N3 + 1.578 \cdot Nx \end{aligned} \quad (18.6)$$

TABLE 18.1
Factors Analysed in the Lung Cancer Study

Factor	Coding (range)	Interpretation	Median or # Cases
Age	30–75	Age at onset of therapy	60
Ecog	1: ecog = 0 or 1	Performance status	558
	2: ecog = 2		43
Stage	1: limited disease	Stage of disease	250
	2: extended disease		351
Tumor	1: T1	Tumor size (TNM classification)	18
	2: T2		158
	3: T3		94
	4: T4		328
	5: Tx		3
Lymph nodes	0: N0	Status of lymph node affection (TNM classification)	22
	1: N1		30
	2: N2		205
	3: N3		314
	4: Nx		30
Metastasis	0: M0	Status of metastasis (TNM classification)	271
	1: M1		301
	2: Mx		29
Gender	1: female	Gender of the patient	148
	2: male		453

TABLE 18.2
Result of the Use of Fractional Polynomials for the Factor Stage

Factor	$\beta(t)$	$p(H_0: \beta(t) = \beta)$	$p(H_0: \beta(t) = 0)$
Stage	$1.005 \cdot t - 0.213 \cdot t^2$	0.0089	<0.001

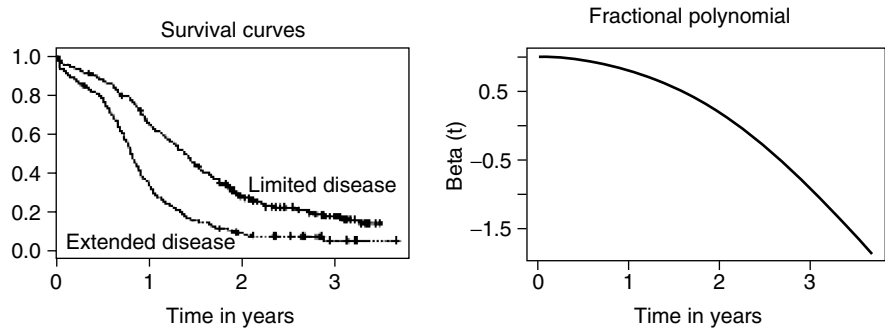


FIGURE 18.2 Results of the analysis for the variable stage.

18.3.3 BUMP HUNTING (IDENTIFICATION OF PREDICTIVE FACTORS)

Now we applied the prognostic model (18.6) together with the corresponding base-line hazard rate $\Lambda_0(t)$ to the patients under the new treatment TEC. Thus we gain martingale residuals shown in Figure 18.3, which will be used as response variables for the bump hunting process stabilized with bootstrapping.

We first searched for patients who are harmed by the new treatment, e.g., who have positive martingale residuals. At each step of the border selection process we compared the remaining patients in the two treatment arms by the log rank test. We stopped the procedure when there appeared a significant difference in survival. The algorithm found one box consisting of four borders. Table 18.3 shows the involved borders, the number of remaining patients, and the p -value of the log rank test for each border.

Only the patients situated in the final bump are considered as being harmed by the new therapy and do better under the classical treatment CEV. These patients are females, have a lymph node affection of N2 or less, a tumor size of T3 or higher, and

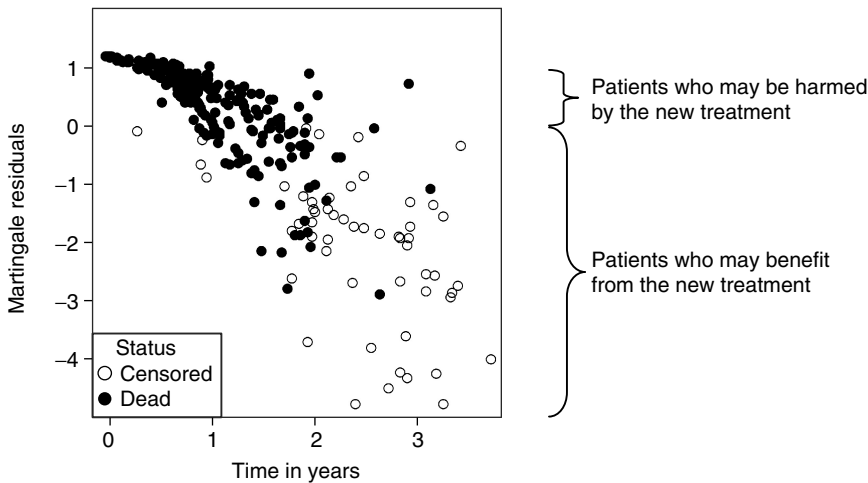


FIGURE 18.3 Martingale residuals of the patients treated with TEC.

TABLE 18.3
Result of Bump Hunting: Selection of Borders and the Corresponding Data

No. of Border	Border	No. of Patients		Mean of Residuals		No. of Events		Log rank Test p(LR)
		TEC	CEV	TEC	CEV	TEC	CEV	
Whole data		296	305	−0.2726	0	240	269	0.0210
Border 1	Gender = 1	75	73	−0.0316	−0.0043	60	59	0.6355
Border 2	Lymph nodes ≤ 2	29	30	0.1294	0.0503	24	24	0.5717
Border 3	Tumor = T3 or T4	17	17	0.4543	−0.1174	15	13	0.2509
Border 4	Metastasis = M1	8	9	0.6982	−0.1814	8	8	0.0026

possess distant metastases. This group contains only 3% of all patients. Figure 18.4 displays the survival curves of the patients being harmed by the new treatment TEC and the survival curves of the rest of the patients.

18.4 FINAL REMARKS

The identification of predictive factors is of great interest in medicine. Clinicians want to know which therapy will be effective in a particular patient. There are several proposals in the literature for how to select these factors. The approaches are mostly based on regression models including interaction terms between the treatment and certain factors. The analysis is usually restricted to first order interactions.

We are describing an alternative method based on bump hunting. With this method it is possible to detect interactions of higher order. First it is important to distinguish between prognostic and predictive factors. Prognostic factors have an influence on the

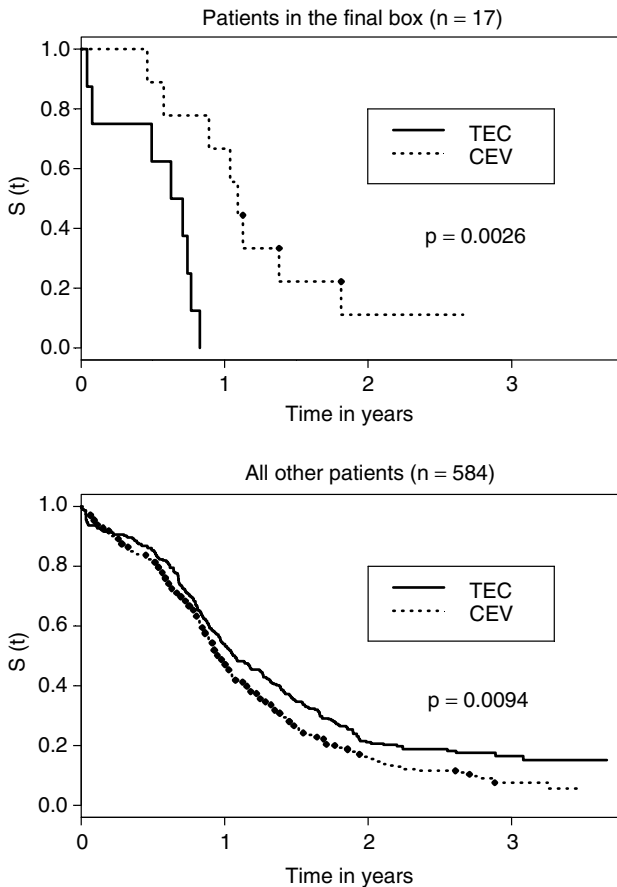


FIGURE 18.4 Survival curves of the patients in the final box and the rest of the data for both treatment groups ($p(\text{LR}) = 0.0026$ and 0.0094 , respectively).

prognosis independent of the therapy. Predictive factors are related to the outcome only under one treatment. In order to distinguish between both groups of factors, a prognostic model has to be established. We used the patients under the standard treatment or from the placebo group. To perform this method, it is necessary to build a good prognostic model. Thus, the extension of the classical Cox proportional hazards model with fractional polynomials is a practicable feature to improve the goodness of fit. However other models, like CART, can also be used.

The prognostic model is then applied to the patients under the new therapy. Martingale residuals are calculated to describe the deviation from the prognostic model. We used bump hunting to look for similarities of patients not well described by the prognostic model. The term *prognostic model* is used. However, if this model is derived from patients receiving the standard treatment, the factors selected may be predictive for this particular therapy. Nevertheless, we used the term *prognostic model*.

Alternatively CART could be applied. Both methods, bump hunting and CART, have advantages and disadvantages. Within CART a split is always a “hard” decision. The decision in bump hunting is soft. First the sample is reduced by a small amount, e.g., 5% of the data, and second the remaining data are considered for further separation.

In the study analyzed, the new therapy was overall superior to the standard regimen. However, it was possible to identify a small subgroup of patients who seem to be harmed by the new therapy.

One should keep in mind that all the analyses proposed are *post hoc* considerations, and the results might be due to chance. Therefore it is “the best advice to report them sceptically as hypotheses to be investigated in other studies”.²¹ Assmann et al.²² similarly state that investigators should be cautious when undertaking subgroup analysis. The findings are always exploratory and one should avoid overinterpretation.

REFERENCES

1. Isaacs, C., Stearns, V., and Hayes, D.F., New prognostic factors for breast cancer recurrence, *Seminars in Oncology*, 28, 53–67, 2001.
2. Look, M. and Ulm, K., Pooled analysis of prognostic impact of uPA and PAI-I in breast cancer patients, *Thrombosis and Haemostasis*, 90, 538–548, 2003.
3. Royston, P. and Sauerbrei, W., A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials, *Statistics in Medicine*, 23, 2509–2525, 2004.
4. Bonetti, M. and Gelber, R.D., Patterns of treatment effects in subsets of patients in clinical trials, *Biostatistics*, 5, 465–481, 2004.
5. Ciampi, A., Negassa, A., and Lou, Z., Tree-structured prediction of censored survival data and the Cox model, *Journal of Clinical Epidemiology*, 48, 675–689, 1995.
6. Breiman, L. et al., *Classification and Regression Trees*, Wadsworth, Belmont, 1984.
7. O’Quigley, J. and Natarajan, L., Erosion of regression effect in a survival study, *Biometrics*, 60, 344–351, 2004.
8. Cox, D.R., Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 187–220, 1972.

9. Hastie, T.J. and Tibshirani, R., *Generalized Additive Models*, Chapman & Hall, London, 1990.
10. Green, P. and Silverman, B., *Nonparametric Regression and Generalized Linear Models*, Chapman & Hall, London, 1994.
11. Royston, P. and Altman, D.G., Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling, *Applied Statistics*, 43, 429–467, 1994.
12. Berger, U., Ulm, K., and Schäfer, J., Dynamic Cox modelling based on fractional polynomials: time-variation in gastric cancer prognosis, *Statistics in Medicine*, 22, 1163–1180, 2003.
13. Friedman, J. and Fisher, N.I., Bump hunting in high-dimensional data, *Statistics and Computing*, 9.2, 123–143, 1999.
14. Barlow, W.E. and Prentice, R.L., Residuals for relative risk regression, *Biometrika*, 75, 65–74, 1988.
15. Gill, R.D., Understanding Cox's regression model: a martingale approach, *Journal of the American Statistical Association*, 79, 441–447, 1984.
16. Therneau, T.M., Grambsch, P.M., and Fleming, T.R., Martingale hazards, regression models, and the analysis of censored survival data, *Biometrika*, 77, 147–160, 1990.
17. Kaplan, E.L. and Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457–481, 1958.
18. Peto, R. and Peto, J., Asymptotically efficient rank invariant test procedures, *Journal of the Royal Statistical Society, Series A*, 135, 185–198, 1972.
19. Kehl, V. Responder Identification in Clinical Trials, Ph.D. thesis, Institut für Mathematik, Informatik und Statistik, Ludwig-Maximilians-Universität, München, online publication, http://edoc.ub.uni-muenchen.de/archive/00000590/Kehl_Victoria.pdf, 2002.
20. Reck, M. et al. Randomized phase III trial of Paclitaxel, Etoposide, and Carboplatin versus Carboplatin, Etoposide, and Vincristin in patients with small-cell lung cancer, *Journal of the National Cancer Institute*, 95, 1118–1127, 2003.
21. Byar, D., Assessing apparent treatment: covariate interactions in randomized clinical trials, *Statistics in Medicine*, 4, 255–263, 1985.
22. Assmann, S.F. et al., Subgroup analysis and other (mis)uses of baseline data in clinical trials, *The Lancet*, 355, 1064–1068, 2000.

19 Explained Variation in Proportional Hazards Regression

John O'Quigley and Ronghui Xu

CONTENTS

19.1	Introduction	347
19.1.1	Motivation	347
19.1.2	A Brief Overview	348
19.2	Model and R^2 Measure	350
19.2.1	Model and Notation	350
19.2.2	Measure of Explained Variation	351
19.2.2.1	The Measure R^2	351
19.2.2.2	Population Parameter Ω^2	352
19.3	Properties and Interpretation	354
19.3.1	Properties of R^2 and Ω^2	354
19.3.2	Interpretation	355
19.4	Extensions	356
19.4.1	Partial Coefficients	356
19.4.2	Stratified Model	356
19.4.3	Other Relative Risk Models	357
19.5	Applications in Cancer Studies	358
19.5.1	Example 1	358
19.5.2	Example 2	358
19.5.3	Example 3	359
19.5.4	Example 4	360
	Appendix	362
	References	362

19.1 INTRODUCTION

19.1.1 MOTIVATION

In a study of 2174 breast cancer patients followed over a period of 15 years at the Institut Curie in Paris, France, a large number of potential and known prognostic

factors were recorded. Detailed analyses of these data have been the subject of a number of communications, and we focus here on a limited analysis on a subset of prognostic factors identified as having some prognostic importance. These factors were: (1) age at diagnosis, (2) histology grade, (3) stage, (4) progesterone receptor status, and (5) tumor size. In addition to the usual model fitting and diagnostic tools, it seems desirable to be able to present summary measures estimating the percentage of explained variation. Making precise the notion of explained variation in the context of proportional hazards regression, in which we allow inference to be invariant to unspecified monotonically increasing transformations on time, is not immediate, but a suitable measure would reflect the relative importance of the covariates.

We would like to be able to say, for example, that stage explains some 20% of survival but that, once we have taken account of progesterone status, age, and grade, then this figure drops to 5%. Or that after adding tumor size to a model in which the main prognostic factors are already included, then the explained variation increases, say, a negligible amount, specifically from 32% to 33%. Or, given that a suitable variable indicates predictability, then to what extent do we lose (or gain), in terms of these percentages, by recoding the continuous prognostic variable, age at diagnosis, into discrete classes on the basis of cutpoints.

For a measure to be able to deal with the above requirements, it would be required to have certain properties: the above percentages should be meaningful and directly related to predictability of the failure ranks; absence of effect should translate as 0%; perfect prediction of the survival ranks should translate as 100%; and intermediate values should be interpretable. The measure introduced by O'Quigley and Flandre (1994), viewable as an index of prediction, a correlation measure, or a measure of explained variation, satisfies these properties. However, in order to be able to recommend the measure for general use, we establish further properties, both statistical and interpretative. The measure can be viewed as a measure of explained variation. One interesting and practically useful property is the ability to accommodate time dependent covariates. We study more deeply the motivation and interpretation of the index as well as showing, via practical examples, how the R^2 index can throw light on the complex interrelations between the risk factors and prognosis in cancer studies. The index can be shown to: (1) take a value of zero in the absence of regression effect, (2) increase to the value 1 when regression effect tends to infinity and prediction of the ranks becomes deterministic, (3) have the property that increasing values of its population counterpart Ω^2 correspond to increasing predictability of the ranks of the survival times, (4) remain invariant to linear transformation of the covariates and to increasing monotonic transformations of time, and (5) enjoy a concrete interpretation in terms of sums of squares decompositions and the ability to express Ω^2 as a proportion of explained variation. Furthermore, Ω^2 can be shown not to be affected by an independent censorship. Extension to the stratified proportional hazards model and to other relative risk models is straightforward.

19.1.2 A BRIEF OVERVIEW

For the proportional hazards model, correlation measures were first suggested by Maddala (1983), although the measure depends heavily on censoring. Kent and

O'Quigley (1988) developed a measure based on the Kullback–Leibler information gain, and this could be interpreted as the proportion of randomness explained in the observed survival times by the covariates. The principle difficulty in Kent and O'Quigley's measure was its complexity of calculation, although a very simple approximation was suggested and appeared to work well. The Kent and O'Quigley measure was not able to accommodate time-dependent covariates. Xu and O'Quigley (1999) developed a similar measure based on information gain using the conditional distribution of the covariates given the failure times. The measure accommodates time-dependent covariates and is computable using standard softwares for fitting the Cox model. Some current work is being carried out on the measures of explained randomness and will be available in a separate manuscript. Schemper (1990, 1994) introduced the concept of individual survival curves for each subject with the model and without the model. Interpretation is difficult. As with the Maddala measure, the Schemper measures depend on censoring, even when the censoring mechanism is completely independent of the failure mechanism (O'Quigley et al., 1999). Korn and Simon (1990) suggested a class of potential functionals of interest, such as the conditional median, and evaluated the explained variation via an appropriate distance measuring the ratio of average dispersions with the model to those without a model. Their measures are not invariant to time transformation, nor could they accommodate time-dependent covariates. Korn and Simon (1990) also suggested an approach to calculate Somer's D; this in general requires numerical integration and depends on a common censoring time for all survival times. Schemper and Kaider (1997) proposed to estimate the correlation coefficient between failure rankings and the covariates via multiply imputing the censored failure times. Other measures have also been proposed in the literature, but it is not our intention here to give a complete review of them.

In the next section we discuss a measure that is a further development of that introduced by O'Quigley and Flandre (1994). We describe the statistical properties of the measure. These properties are particularly attractive from the viewpoint of interpretation, in particular the ability to obtain a sum of squares decomposition analogous to that for the linear case, an expression in terms of explained variation and the observation that increasing values of the measure directly translate increasing predictability of the survival ranks. We discuss the partial coefficients and the extensions of the measure to other relative risk models. An illustration of the practical usefulness of the measure in analyzing survival data under the proportional hazards model is given. This is done via three illustrative examples: breast cancer, gastric cancer, and multiple myeloma. In these studies we are interested in questions such as how much additional prognostic value is contained in some explanatory variable once others have been included in the model, to what extent the observed failure rankings can be explained by the totality of the explanatory variables, to what extent modeling time-dependent effects improves prediction, and to what extent predictive power is diminished, if at all, by recoding a continuous covariate as a discrete dichotomy or trichotomy. The example of multiple myeloma is interesting in that, even using all available molecular and genetic markers, we fail to account for over 80% of the variability in the survival rankings. For this example it is also interesting that biomarkers alone would appear to capture all the predictive capacity of the traditional Durie–Salmon staging system.

19.2 MODEL AND R^2 MEASURE

19.2.1 MODEL AND NOTATION

In a survival study, denote T the potential failure time and C the potential censoring time. Let $X = \min(T, C)$, $\delta = I(T \leq C)$ where $I(\cdot)$ is the indicator function, and $Y(t) = I(X \geq t)$. Associated with T is the vector of possibly time-dependent covariates $Z(t)$. For our mathematical development, assume $(T_i, C_i, Z_i(\cdot))$, $i = 1, 2, \dots, n$, to be a random sample from the distribution of $(T, C, Z(\cdot))$. We will also use the counting process notation: let $N_i(t) = I\{T_i \leq t, T_i \leq C_i\}$ and $\bar{N}(t) = \sum_1^n N_i(t)$. For most of this work we assume a conditional independent censorship model. However, under the stronger assumption of independent censorship, where C is assumed to be independent of T and Z , we obtain further properties and interpretation as described in Section 19.3.

The Cox (1972) proportional hazards model assumes that the conditional hazard function

$$\lambda(t|Z(t)) = \lambda_0(t) \exp\{\beta'Z(t)\} \quad (19.1)$$

where $\lambda_0(t)$ is an unknown “baseline” hazard and β is the relative risk parameter. Denote

$$\pi_i(\beta, t) = \frac{Y_i(t) \exp\{\beta'Z_i(t)\}}{\sum_{j=1}^n Y_j(t) \exp\{\beta'Z_j(t)\}} \quad (19.2)$$

the conditional probability of subject i being chosen to fail, given all the individuals at risk at time t and that one failure occurs. The product of the π 's over the observed failure times gives Cox's (1972, 1975) partial likelihood. When $\beta = 0$, $\{\pi_i(0, t)\}_{i=1}^n$ is simply the empirical distribution, assigning equal weight to each sample subject in the risk set. Denote the expectation of a variable with respect to $\{\pi_i(\beta, t)\}_{i=1}^n$ by $\varepsilon_\beta(\cdot|t)$. In particular

$$\varepsilon_\beta(Z|t) = \sum_{j=1}^n Z_j(t) \pi_j(\beta, t) = \frac{\sum_{j=1}^n Y_j(t) Z_j(t) \exp\{\beta'Z_j(t)\}}{\sum_{j=1}^n Y_j(t) \exp\{\beta'Z_j(t)\}} \quad (19.3)$$

is the expectation of $Z(t)$ with respect to $\{\pi_i(\beta, t)\}_i$, and

$$r_i(\hat{\beta}) = Z_i(X_i) - \varepsilon_\beta(Z|X_i) \quad (19.4)$$

for $\delta_i = 1$ is the Schoenfeld (1982) residual where $\hat{\beta}$ is usually obtained by solving the estimating equation provided by the first derivative of the log-partial-likelihood, $U(\beta) = 0$. It is interesting to note that $U(\beta) = \sum_{i=1}^n \delta_i r_i(\beta)$. Analogous to ordinary regression, we make the sum of the residuals equal to zero in order to estimate the unknown parameter. We consider the sum of squared residuals to study predictability.

19.2.2 MEASURE OF EXPLAINED VARIATION

19.2.2.1 The Measure R^2

Let us first assume Z of dimension one. In (19.3) the expectation $\varepsilon_\beta(Z|X_i)$ is worked out with respect to an exponentially tilted distribution. The stronger the regression effects the greater the tilting, and the smaller we might expect, on average, the values $r_i^2(\beta)$ to be when compared with the residuals under the null model $\beta = 0$. Based on these residuals, a measure of explained variation can be defined (O'Quigley and Flandre, 1994), which is analogous to the coefficient of determination for the linear model. In the absence of censoring, the quantity $\sum_{i=1}^n r_i^2(\hat{\beta})/n$ is a residual sum of squares and can be viewed as the average discrepancy between the observed covariate and its expected value under the model, whereas $\sum_{i=1}^n r_i^2(0)/n$ is a total sum of squares and can be viewed as the average discrepancy without a model. Because the semiparametric model leaves inference depending only on the failure time rankings, and being able to predict failure rankings of all the failed subjects is equivalent to being able to predict at each failure time that subject is to fail, it is sensible to measure the discrepancy between the observed covariate at a given failure time and its expected value under the model. Thus we can define $\mathcal{I}(b)$ for $b = 0, \beta$ by

$$\mathcal{I}(b) = \sum_{i=1}^n \delta_i r_i^2(b) \quad (19.5)$$

so that $\mathcal{I}(\hat{\beta}) = \sum_{i=1}^n \delta_i r_i^2(\hat{\beta})$, the sum of the squared residuals whereas $\mathcal{I}(0) = \sum_{i=1}^n \delta_i r_i^2(0)$, is the sum of the squared “null” residuals, or the total sum of squares (made more precise below). We then define

$$R^2(\beta) = 1 - \frac{\mathcal{I}(\beta)}{\mathcal{I}(0)} \quad (19.6)$$

which corresponds to the definition given by O'Quigley and Flandre (1994). For a normal model, considering the residuals in the above formula to be the usual normal residuals, then this definition coincides exactly with the usual coefficient of correlation, also interpretable as a percentage of explained variation. Generalizing that definition is then very natural. For the multivariate case, when $Z(t)$ is a $p \times 1$ vector, the dependence of the survival time variable on the covariates is best expressed via the prognostic index (Andersen et al., 1983; Altman and Andersen, 1986)

$$\eta(t) = \beta' Z(t)$$

Two individuals with possibly different Z s but the same η will have the same survival probabilities. So we can imagine that each subject in the study is now labeled by η . R^2 as a measure of explained variation or, predictive capability, should evaluate how well the model predicts which individual or equivalently, its label, is chosen to fail at each observed failure time. This is equivalent to predicting the failure rankings given the prognostic indices. When $p = 1$, Z is equivalent to η ; therefore, we can construct the R^2 using residuals of the Z s. But for $p > 1$, the model does not distinguish between

different vector Z s as long as the corresponding η s are the same. So instead of residuals of Z , we define the multiple coefficient using residuals of η . Analogous to the univariate definition, we have, once again, $R^2(\beta) = 1 - \mathcal{I}(\beta)/\mathcal{I}(0)$ where, for the multivariate case

$$\mathcal{I}(b) = \sum_{i=1}^n \delta_i \{\beta' r_i(b)\}^2 \quad (19.7)$$

19.2.2.2 Population Parameter Ω^2

The population parameter $\Omega^2(\beta)$ of $R^2(\hat{\beta})$ was originally given in O'Quigley and Flandre (1994). However, as discussed in Xu (1996), in order to completely eliminate any asymptotic dependence upon censoring, it is necessary to weight the squared Schoenfeld residuals by the increments of any consistent estimate of the marginal failure time distribution function $F(t)$. The survivorship function is just $S(t) = 1 - F(t)$. The practical impact of this weighting on numerical values would typically be small and, in routine analysis, we might choose to work with the simpler calculation. Our main purpose in giving consideration to this weighting of the residuals is to provide a mathematically tight framework to the large sample theory. Therefore, let $\hat{F}$ be the left-continuous Kaplan–Meier (KM) estimate of F , and define $W(t) = \hat{S}(t)/\sum_{i=1}^n Y_i(t)$ where $\hat{S} = 1 - \hat{F}$. Then $W(t)$ is a nonnegative predictable stochastic process and, assuming there are no ties, it is straightforward to verify that $W(X_i) = \hat{F}(X_i+) - \hat{F}(X_i)$ at each observed failure time X_i , i.e., the jump of the KM curve. In practice, ties, if they exist, are split randomly. So in place of the above definition of $\mathcal{I}(b)$ for $b = 0, \beta$, we use a more general definition in which

$$\mathcal{I}(b) = \sum_{i=1}^n \int_0^\infty W(t) \{Z_i(t) - \varepsilon_b(Z|t)\}^2 dN_i(t) = \sum_{i=1}^n \delta_i W(X_i) r_i^2(b) \quad (19.8)$$

We now define

$$R^2(\beta) = 1 - \frac{\sum_{i=1}^n \delta_i W(X_i) r_i^2(\beta)}{\sum_{i=1}^n \delta_i W(X_i) r_i^2(0)} = 1 - \frac{\mathcal{I}(\beta)}{\mathcal{I}(0)} \quad (19.9)$$

The definition given by O'Quigley and Flandre (1994) would be the same as above if we defined $W(t)$ to be constant and, of course, the two definitions coincide in the absence of censoring. The motivation for the introduction of the weight $W(t)$ is to obtain large sample properties of R^2 that are unaffected by an independent censoring mechanism. Viewing R^2 as a function of β turns out to be useful in theoretical studies. In practice, we are mostly interested in $R^2(\hat{\beta})$ where $\hat{\beta}$ is a consistent estimate of β such as the partial likelihood estimate. Let

$$S^{(r)}(\beta, t) = n^{-1} \sum_{i=1}^n Y_i(t) e^{\beta' Z_i(t)} Z_i(t)^{\otimes r}, \quad s^{(r)}(\beta, t) = ES^{(r)}(\beta, t) \quad (19.10)$$

for $r = 0, 1, 2$. Here $a^{\otimes 2} = aa'$ and $a \otimes b = ab'$ for vectors a and b , $a^{\otimes 1} = a$ and $a^{\otimes 0} = 1$. We assume that conditions **A–C** of the [appendix](#) hold. Notice that $\varepsilon_\beta(Z|t) = S^{(1)}(\beta, t)/S^{(0)}(\beta, t)$. Let

$$J(\beta, b) = \int w(t) \beta' \left\{ \frac{s^{(2)}(\beta, t)}{s^{(0)}(\beta, t)} - 2 \frac{s^{(1)}(\beta, t) \otimes s^{(1)}(b, t)}{s^{(0)}(\beta, t) s^{(0)}(b, t)} + \frac{s^{(1)}(b, t)^{\otimes 2}}{s^{(0)}(b, t)^2} \right\} \beta s^{(0)}(\beta, t) \lambda_0(t) dt \quad (19.11)$$

where $w(t) = S(t)/s^{(0)}(0, t)$. Then

$$\Omega^2(\beta) = 1 - \frac{J(\beta, \beta)}{J(\beta, 0)} \quad (19.12)$$

Notice that although (19.6) and (19.12) are not defined for $\beta=0$, the limits exist and are equal to zero as $\beta \rightarrow 0$. So we can define $R^2(0) = \Omega^2(0) = 0$.

It has been shown (Xu, 1996) that $\Omega^2(\beta)$ is unaffected by an independent censorship mechanism, i.e., when C is independent of T and Z , and in this case it can be written (O'Quigley and Flandre, 1994)

$$\Omega^2(\beta) = 1 - \frac{\int E_\beta\{[Z(t) - E_\beta(Z(t)|t)]^2|t\} dF(t)}{\int E_\beta\{[Z(t) - E_0(Z(t)|t)]^2|t\} dF(t)} \quad (19.13)$$

If, in addition, Z is time-invariant, we will see that $\Omega^2(\beta)$ has the interpretation of the proportion of explained variation (Section 19.3.2). O'Quigley and Flandre showed that, having standardized for the mean and the variance, $\Omega^2(\beta)$ depends only relatively weakly on different covariate distributions, and values of $\Omega^2(\beta)$ appear to give a good reflection of strength of association as measured by β and tend to 1 for high but plausible values of β . Their numerical results support the conjecture that Ω^2 increases with the strength of effect, thereby agreeing with the third stipulation of Kendall (1975, p. 4) for a measure of rank correlation. The conjecture was proven to be true in Xu (1996); see also Section 19.3.1. The first two stipulations were that perfect agreement or disagreement should reflect itself in a coefficient of absolute value 1; the third stipulation was that for other cases the coefficient should have absolute value less than 1, and that in some acceptable sense, increasing values of the coefficient should correspond to increasing agreement between the ranks. Here we have a squared coefficient, and Kendall's stipulations are considered in a broader sense because we are not restricted to the ranks of the covariates in the semiparametric context.

19.3 PROPERTIES AND INTERPRETATION

In this section we show that the measure defined above has the desired properties and the interpretation as a measure of explained variation. We omit all the proofs here. They can be found in Xu (1996).

19.3.1 PROPERTIES OF R^2 AND Ω^2

The R^2 defined above can be shown to have the following properties:

1. $R^2(0) = 0$.
2. $R^2(\hat{\beta}) \leq 1$.
3. $R^2(\beta)$ is invariant under linear transformations of Z and monotonically increasing transformations of T .
4. $R^2(\hat{\beta})$ consistently estimates $\Omega^2(\beta)$. In particular, $\mathcal{I}(\hat{\beta})$ and $\mathcal{I}(0)$ consistently estimate $J(\beta, \beta)$ and $J(\beta, 0)$, respectively.
5. $R^2(\beta)$ is asymptotically normal.

Note that in finite samples, R^2 , unlike Ω^2 below, cannot be guaranteed to be non-negative. A negative value for R^2 would correspond to the unusual case in which the best fitting model, in a least squares sense, provides a poorer fit than the null model. Our experience is that $R^2(\hat{\beta})$ will only be slightly negative in finite samples if $\hat{\beta}$ is very close to zero.

Similarly, we have the following properties for Ω^2 :

1. $\Omega^2(0) = 0$.
2. $0 \leq \Omega^2(\beta) \leq 1$.
3. $\Omega^2(\beta)$ is invariant under linear transformations of Z and monotonically increasing transformations of T .
4. For a scalar β , $\Omega^2(\beta)$ as a function of β , increases with $|\beta|$; and as $|\beta| \rightarrow \infty$, $\Omega^2(\beta) \rightarrow 1$.

From the last property above, one can show that Ω^2 increases with the predictability of survival rankings, i.e., $P(T_i > T_j)$ for given Z_i and Z_j (assuming without loss of generality that $\beta > 0$). This corresponds to Kendall's third stipulation in the context of semiparametric Cox regression. As result of the last property, we can obtain confidence intervals of $\Omega^2(\beta)$ from those for β because $\Omega^2(\beta)$, viewed as a function of β , is an increasing function of $|\beta_0|$. The coverage properties will then be the same as those already established for the log relative-risk estimate. $R^2_\epsilon(\beta)$ is consistent for $\Omega^2(\beta)$ for any β , so in practice we only need to "plug" the two endpoints of the β -confidence interval into R^2_ϵ . This leads to a confidence interval for $\Omega^2(\beta)$. An alternative approach is via the bootstrap. Following Efron (1981) we can re-sample the triplets (X_i, Z_i, δ_i) to obtain the bootstrap distribution. Making corrections for bias and transforming, via an acceleration adjustment, improve the normality approximation (Hinkley, 1988). The two approaches can be seen to give good agreement in the examples.

19.3.2 INTERPRETATION

In order to be completely assured before using R^2 in practice, it is important to know that R^2 is consistent for Ω^2 , that $\Omega^2(0) = R^2(0) = 0$, $\Omega^2(\infty) = 1$, that Ω^2 increases as strength of effect increases, and that Ω^2 is unaffected by an independent censoring mechanism. This enables us to state that an Ω^2 of 0.4 translates greater predictability than an Ω^2 of 0.3. We do, however, need one more thing. We would like to be able to say precisely just what a value such as 0.4 corresponds to. That is the purpose of this subsection. In definition (19.6) of $R^2(\beta)$, $\sum_{i=1}^n \delta_i W(X_i) \{\hat{\beta}' r_i(\beta)\}^2$ can be considered as a residual sum of squares analogous to the linear regression case, while $\sum_{i=1}^n \delta_i W(X_i) \{\hat{\beta}' r_i(0)\}^2$ is the total sum of squares.

So define

$$SS_{\text{tot}} = \sum_{i=1}^n \delta_i W(X_i) \{\hat{\beta}' r_i(0)\}^2$$

$$SS_{\text{res}} = \sum_{i=1}^n \delta_i W(X_i) \{\hat{\beta}' r_i(\hat{\beta})\}^2$$

$$SS_{\text{reg}} = \sum_{i=1}^n \delta_i W(X_i) \{\hat{\beta}' \varepsilon_\beta(Z|X_i) - \hat{\beta}' \varepsilon_0(Z|X_i)\}^2$$

It can be shown that an asymptotic decomposition holds of the total sum of squares into the residual sum of squares and the regression sum of squares, i.e.,

$$SS_{\text{tot}} \doteq SS_{\text{res}} + SS_{\text{reg}} \quad (19.14)$$

where $\doteq$ means equality in a large sample sense, i.e., the difference between the two sides of the equation converge to zero in probability as $n \rightarrow \infty$. So R^2 is asymptotically equivalent to the ratio of the regression sum of squares to the total sum of squares. For time-invariant covariates and independent censoring, the coefficient $\Omega^2(\beta)$ has a simple interpretation in terms of explained variation, i.e.,

$$\Omega^2(\beta) \approx 1 - \frac{E\{\text{Var}(Z|T)\}}{\text{Var}(Z)} = \frac{\text{Var}\{E(Z|T)\}}{\text{Var}(Z)} \quad (19.15)$$

Indeed there is nothing to stop us defining explained variation as in the right hand side of (19.15) because the marginal distribution of Z and T can be estimated by the empirical and the KM estimator, while the conditional distribution of Z given $T = t$ by the $\{\pi_i(\hat{\beta}, t)\}_i$. However, we can see no advantage to this and recommend that all calculations be done via the Schoenfeld residuals evaluated at $\beta = \hat{\beta}$ and $\beta = 0$.

19.4 EXTENSIONS

19.4.1 PARTIAL COEFFICIENTS

The partial coefficient can be defined via a ratio of multiple coefficients of different orders. Specifically, and in an obvious change of notation just for the purposes of this subsection, let $R^2(Z_1, \dots, Z_p)$ and $R^2(Z_1, \dots, Z_q)$ ($q < p$) denote the multiple coefficients with covariates Z_1 to Z_p and covariates Z_1 to Z_q , respectively. Note that $R^2(Z_1, \dots, Z_p)$ is calculated using $\hat{\beta}_1, \dots, \hat{\beta}_p$ estimated when $Z_1, \dots, Z_p$ are included in the model, and $R^2(Z_1, \dots, Z_q)$ using $\hat{\beta}_{10}, \dots, \hat{\beta}_{q0}$ estimated when only $Z_1, \dots, Z_q$ are included. Define the partial coefficient $R^2(Z_{q+1}, \dots, Z_p | Z_1, \dots, Z_q)$, the correlation after having accounted for the effects of Z_1 to Z_q by

$$1 - R^2(Z_1, \dots, Z_p) = [1 - R^2(Z_1, \dots, Z_q)][1 - R^2(Z_{q+1}, \dots, Z_p | Z_1, \dots, Z_q)] \quad (19.16)$$

The above coefficient, motivated by an analogous expression for the multivariate normal model, makes intuitive sense in that the value of the partial coefficient increases as the difference between the multiple coefficients increases and takes the value zero should this difference be zero. Partial Ω^2 can be defined in a similar way.

We can also derive definition (19.16) directly. Following the discussion of multiple coefficients, we can use the prognostic indices obtained under the model with $Z_1, \dots, Z_p$ and that with $Z_1, \dots, Z_q$. This would be equivalent to defining $1 - R^2(Z_{q+1}, \dots, Z_p | Z_1, \dots, Z_q)$ as $\mathcal{I}(Z_1, \dots, Z_p) / \mathcal{I}(Z_1, \dots, Z_q)$, the ratio of the numerators of $1 - R^2(Z_1, \dots, Z_p)$ and $1 - R^2(Z_1, \dots, Z_q)$. However, because the two numerators are on different scales, being inner products of vectors of different dimensions, their numerical values require standardization. One natural way to standardize is to divide these numerators by the denominators of $1 - R^2(Z_1, \dots, Z_p)$ and $1 - R^2(Z_1, \dots, Z_q)$, respectively. This gives definition (19.16).

Partial coefficients in O'Quigley and Flandre (1994) were defined using a single component of the covariate vector instead of the prognostic index. Although our limited data experience did not show any important discrepancies between that definition and (19.16), there seems to be some arbitrariness as to which component of the vector to use. Furthermore the prognostic index should reflect the best prediction a given model can achieve in the sense we described before. Our recommendation is to use (19.16) as the partial coefficient.

19.4.2 STRATIFIED MODEL

The partial coefficients of the previous section enable us to assess the impact of one or more covariates while adjusting for the effects of others. This is carried out in the context of the assumed model. It may sometimes be preferable to make weaker assumptions than the full model and adjust for the effects of other multilevel covariates by stratification. Indeed it can be interesting and informative to compare adjusted R^2 measures, the adjustments having been made either via the model or via stratification. For the stratified model the definitions of Section 19.2 follow through readily. To be precise, we define a stratum specific residual for stratum s ($s = 1, \dots, S$), where,

in the following, a subscript is in place of i means the i th subject in stratum s . Thus we have

$$r_i(b; s) = Z_{is}(X_{is}) - \varepsilon_b(Z|X_{is}) \quad (19.17)$$

where $\varepsilon_b(Z|X_{is})$ is averaged within stratum s over the risk set at time X_{is} , and we write

$$\mathcal{I}(b) = \sum_i \sum_s \int_0^\infty W(t) \{Z_{is}(t) - \varepsilon_b(Z|t)\}^2 dN_{is}(t) = \sum_i \sum_s \delta_{is} W(X_{is}) r_i^2(b, s) \quad (19.18)$$

From this we can define

$$R^2(\beta) = 1 - \frac{\sum_i \sum_s \delta_{is} W(X_{is}) r_i^2(\beta, s)}{\sum_i \sum_s \delta_{is} W(X_{is}) r_i^2(0, s)} = 1 - \frac{\mathcal{I}(\beta)}{\mathcal{I}(0)} \quad (19.19)$$

Note that we do not use a stratum specific $W(t)$ and, as before, we work with an assumption of a common underlying marginal survival distribution. The validity of this hinges upon an independent rather than a conditionally independent censoring mechanism. Under a conditionally independent censoring mechanism, a weighted Kaplan–Meier estimate (Murray and Tsiatis, 1996) of the marginal survival distribution may be used instead.

19.4.3 OTHER RELATIVE RISK MODELS

It is straightforward to generalize the R^2 measure to other relative risk models, with the relative risk of forms such as $1 + \beta z$ or $\exp\{\beta(t)z\}$. Denote $r(t; z)$ a general form of the relative risk. Assume that the regression parameters involved have been estimated, and define $\pi_i(t) = Y_i(t)\hat{r}(t; Z_i)/\sum_{j=1}^n Y_j(t)\hat{r}(t; Z_j)$. Then we can similarly define $\varepsilon_\beta(Z|t)$ and form the residuals, thereby defining an R^2 measure similar to (19.9). In addition, it can be shown that under an independent censorship, the conditional distribution of $Z(t)$ given $T=t$ is consistently estimated by $\{\pi_i(t)\}_i$, so properties such as being unaffected by an independent censorship are maintained.

It is particularly interesting to study the use of such an R^2 measure under the time-varying regression effects model, where the relative risk is $\exp\{\beta(t)z\}$. Different approaches have been proposed to estimate $\beta(t)$ (Sleeper and Harrington, 1990; Zucker and Karr, 1990; Murphy and Sen, 1991; Gray, 1992; Hastie and Tibshirani, 1993; Verweij and Van Houwelingen, 1995; Sargent, 1997; Gustafson, 1998; Xu and Adak, 2001). In this case we can use R^2 to compare the predictability of different covariates as we do under the proportional hazards model; we can also use it to guide the choice of the amount of smoothness, or the “effective degrees of freedom” as it is called by some of the aforementioned authors, in estimating $\beta(t)$. As a brief illustration, suppose that we estimate $\beta(t)$ as a step function and that we are to choose between two different partitions of the time axis, perhaps one finer than the other. Denote the two estimates obtained under these two partitions by $\hat{\beta}_1(t)$ and

$\hat{\beta}_2(t)$, the latter corresponding to the finer partition. We can measure the extra amount of variation explained by fitting $\hat{\beta}_2(t)$ versus fitting $\hat{\beta}_1(t)$ by

$$R_{\text{ex}}^2 = 1 - \frac{\mathcal{I}(\hat{\beta}_2(\cdot))}{\mathcal{I}(\hat{\beta}_1(\cdot))} \quad (19.20)$$

This can be thought of as a partial coefficient if we look at the “dimension” of $\beta(t)$ through time. The use of R_{ex}^2 in estimating $\beta(t)$ was recently adopted in Xu and Adak (2001).

19.5 APPLICATIONS IN CANCER STUDIES

We consider some practical examples illustrating how the R^2 measure can throw light on the complex interrelations between risk factors and prognosis. Here we emphasize the use of R^2 while omitting the detailed data analysis with regard to other aspects of model diagnostics. We used the definition of R^2 incorporating the weights $W(X_i)$. Although not reported here, an analysis without these weights, as described by O’Quigley and Flandre (1994), is for all practical purposes identical.

19.5.1 EXAMPLE 1

The example concerns the well-known Freireich (1963) data, which record the remission times of 42 patients with acute leukemia treated by 6-mercaptopurine (6-MP) or placebo. It was used in the original Cox (1972) paper and has been studied by many other authors under the proportional hazards model. Our estimate of the regression coefficient is $\hat{\beta} = 1.53$, and the weighted value of $R^2(\beta)$ comes out to be 0.386. The unweighted calculation produces the value 0.389. Based on 1000 bootstrap samples, we obtain a simple 95% confidence interval (0.154, 0.714). The bootstrap mean turned out to be 0.413 for R^2 , which gives estimated bias of 0.028. This suggests that bias correction might be necessary, and employing Efron’s bias-corrected accelerated bootstrap (BCa) method, we have a 95% confidence interval (0.111, 0.631) on the basis of R^2 . This value of $R^2(\hat{\beta})$ can be compared with some of the other suggestions mentioned in the introduction. For the same data the measure proposed by Kent and O’Quigley (1988) resulted in the value 0.37, and the similar measure by Xu and O’Quigley (1999) resulted in the value of 0.40. The explained variation proposals of Schemper (1990) based on empirical survival functions per subject resulted in (his notation) $V_1 = 0.20$ and $V_2 = 0.19$, and Schemper’s later correction (1994) resulted in $V_2 = 0.29$, although all three of these measures depend on the censoring (O’Quigley et al. 1999). The measure of Schemper and Kaider (1997) resulted in $r_{pr}^2 = 0.34$. The measure of Korn and Simon (1990) based on quadratic loss gave the value 0.32. This measure does not remain invariant to monotone increasing transformation of time. For these data the value 0.32 drops to 0.29 if the failure times are replaced by the square roots of the times.

19.5.2 EXAMPLE 2

The second illustration concerns the breast cancer study described briefly in the introduction. We first fit different proportional hazard models with a single covariate and cal-

culated the R^2 as defined in (19.6). These results are summarized in Table 19.1. All variables are highly significant. The predictive power, though, is quite different. Stage and tumor size, as one might expect, have reasonably high predictability. Histology grade also has predictive power, although this covariate has been shown to have a nonproportional regression effect. We investigated a more complex model in which the coefficient for histology was allowed to decay with time. The value of R^2 increased from 0.116 to 0.24, the improvement in explained variation reflecting an improvement in fit. This case also underlines the relationship between predictability and goodness-of-fit. On the other hand, age has very weak predictive capability, though significant. This estimated weak effect could be due to: 1) a population weak effect, or 2) a suboptimal coding of the covariate. We investigated this second possibility via two recoded models. The first, making a strong trend assumption, coded age as 1(0–33), 2 (34–40) and 3 (41 and above). The second model, making no assumptions about trend, used two binary variables to code the three groups. All three models gave very similar values of R^2 . In consequence only the simplest model is retained for subsequent analysis, i.e., the age groups 1–3. In addition, we calculated the multiple R^2 for a set of nested models. These results are illustrated in Table 19.2. Table 19.2 also contains the values of the partial R^2 defined in (19.16), when each additional covariate is added to the existing model. The partial coefficient for tumor size having accounted for the other four variables is 0.006, suggesting that the extra amount of variation in survival explained by the patient’s tumor size is quite limited. The use of partial R^2 is further explored in Example 3.

19.5.3 EXAMPLE 3

In a study on prognostic factors in gastric cancer (Rashid et al., 1982) certain acute phase reactant proteins were measured preoperatively. Five covariates were studied:

TABLE 19.1
Breast Cancer, Univariate Analysis

Covariate	β	p -Value	R^2
Age	−0.24	<0.01	0.005
Hist	0.37	<0.01	0.12
Stage	0.53	<0.01	0.20
Prog	−0.73	<0.01	0.07
Size	0.02	<0.01	0.18

TABLE 19.2
Breast Cancer, Multivariate Analysis

Covariates	R^2	Partial R^2
Age	0.01	
Age and hist	0.12	0.12
Age, hist and stage	0.26	0.16
Age, hist, stage, and prog	0.33	0.09
Age, hist,stage, prog and size	0.33	0.01

stage together with the proteins α_1 -anti chymotrypsin (ACT), carcino embryonic antigen (CEA), C-reactive protein (CRP), and α_1 glyco-protein (AGP). Surgery is needed in order to determine the stage of the cancer, a clinical factor known to strongly influence survival, and one of the purposes of the study was to find out how well the four protein covariates, available preoperatively, are able to explain survival in the absence of information on stage. A logarithm transformation for CEA was found to be necessary. This is also reflected in a R^2 increasing from 0.10 to 0.20 after the transformation. Table 19.3 shows that each of the five covariates has reasonable predicting power, with R^2 for stage alone to be 0.48. A direct calculation of sample correlation shows that ACT, CRP, and AGP are highly correlated, which is supported by biological evidence. In addition, fitting the Cox model with all four protein covariates shows that CRP and AGP are no longer significant in the presence of the other covariates. These two variables were dropped from further study. The value of R^2 for a model with ACT and log(CEA) is 0.37; this increases to 0.54 when stage is also included, and the corresponding partial R^2 is equal to 0.27. In conclusion, there is strong prognostic information in the preoperative measurements ACT and log(CEA), but this only partially captures the information contained in stage.

19.5.4 EXAMPLE 4

Our fourth example was motivated by the increasing number of correlative studies carried out in cancer research to relate the outcome with multidimensional molecular and genetic markers. The data come from a clinical trial (EST 9486) of multiple myeloma conducted by the Eastern Cooperative Oncology Group (Oken et al., 1999). The trial collected laboratory measurements on patients' myeloma cells, including measurements from the blood or serum (albumin, β_2 microglobulin, creatinine, immunoglobulins IgA and IgG, percent plasma cells, and hemoglobin); characteristics of the circulating myeloma cells (plasma cell labeling index, IL-6 receptor status, and C-reactive protein); and kappa light chain. The study assigned subjects to three randomized treatment arms, but no significant survival difference was found across the three arms. Here we include a randomly selected group of 295 patients, on whom a particular chromosomal abnormality, the possible deletion of the short arm of chromosome 13 (denoted by 13q-), was measured by fluorescent *in situ* hybridization (FISH) in the laboratory of R. Fonseca at the Mayo Clinic. We also include the traditional Durie–Salmon stage, which was routinely used to predict prognosis in multiple myeloma before the availability of assays to measure genetic and other molecular abnormalities of the myeloma cells.

TABLE 19.3
Gastric Cancer, Univariate Analysis

Covariate	$\hat{\beta}$	p-Value	R^2
Stage	1.78	<0.01	0.48
ACT	2.26	<0.01	0.29
Log(CEA)	0.30	<0.01	0.20
CRP	0.02	<0.01	0.26
AGP	0.70	<0.01	0.14

Univariate Cox regression analysis indicates that all of the above 13 covariates are associated with patients' survival times (p -value<0.23), and most of them are highly significant. Table 19.4 shows the estimated regression effects and the standard errors, and the univariate R^2 coefficients. As we see, the predictability by an individual marker is generally low, with the highest R^2 of 0.08 from plasma cell labeling (PCL) index. However, when all the covariates are included in a multiple Cox model, only six of them remain significant (p -value<0.08), with the multiple R^2 =0.20. In particular, the traditional staging system is no longer significantly predictive of survival given the laboratory measurements. Leaving out the nonsignificant variables in a Cox model gives R^2 =0.18. As an illustration of variable selection using R^2 , we build hierarchical models starting with PCL index, which has the highest univariate R^2 . We then choose the variable among the remaining five that has the highest partial R^2 , and so on. Table 19.5 gives the nested models and the corresponding R^2 s.

The above examples serve the purposes of illustration and help underline our arguments that R^2 is useful. They are not in any sense thorough, and in practical applications, deeper study would be useful. For instance, the combinatorial problem of

TABLE 19.4
Myeloma Univariate Analysis

Covariate	$\hat{\beta}$	se ($\hat{\beta}$)	R^2
Creatinine	0.66	0.16	0.05
Plasma	0.43	0.12	0.04
IL-6	0.35	0.14	0.02
C-reactive	0.53	0.17	0.02
a13q	0.22	0.12	0.01
Hemoglobin	−0.30	0.13	0.03
Albumin	−0.39	0.14	0.03
IgG	−0.15	0.12	0.01
IgA	0.16	0.14	0.01
Kappa	−0.26	0.12	0.01
Stage	−0.18	0.12	0.004
β_2 microglobulin	0.48	0.13	0.03
PCL Index	0.59	0.13	0.08

TABLE 19.5
Myeloma Nested Models

Covariates	R^2
PCL	0.08
PCL, creat	0.11
PCL, creat, plasma	0.13
PCL, creat, plasma, a13q	0.16
PCL, creat, plasma, a13q, β_2 mcrglb	0.17
PCL, creat, plasma, a13q, β_2 mcrglb, IL-6	0.18
All 13 Variables	0.20

examining all possible subgroups of different sizes raises both statistical and computational challenges. The statistical question, also present in the limited analyses above, is that of bias, or inflation, of multiple R^2 away from zero when viewed as an estimate of the corresponding multiple Ω^2 . This question is not specific to the survival setting and arises in the standard case of linear regression. Bias reduction techniques used there, such as bootstrap resampling, would also be helpful for our application. In the myeloma example it is quite likely that the small observed increases in R^2 when adding further variables to the vector (PLC, creatinine, plasma, a13q) are artifacts of the data rather than indications that similar increases also hold for Ω^2 .

APPENDIX

For showing large sample properties, we make the following assumptions, which are similar to those in Andersen and Gill (1982):

- A. (Finite interval). $\int_0^1 \lambda_0(t) dt < \infty$.
- B. (Asymptotic stability). There exist a neighborhood $\mathcal{B}$ of β such that 0 and β_0 belong to the interior of $\mathcal{B}$, and

$$\sup_{t \in [0,1]} \|nW(t) - w(t)\| \xrightarrow{P} 0$$

$$\sup_{t \in [0,1], \beta \in \mathcal{B}} \|S^{(r)}(\beta, t) - s^{(r)}(\beta, t)\| \xrightarrow{P} 0$$

for $r = 1, 2, 3, 4$, where the arrows indicate convergence in probability with rate $n^{-1/2}$.

- C. (Asymptotic regularity conditions). All functions in **B** are uniformly continuous in $t \in [0, 1]$; $s^{(r)}(\beta, t)$, $r = 0, 1, 2, 3, 4$, are continuous functions of $\beta \in \mathcal{B}$, and are bounded on $\mathcal{B} \times [0, 1]$; $s^{(0)}(\beta, t)$ is bounded away from zero.

REFERENCES

- Altman, D.G. and Andersen P.K. (1986) A note on the uncertainty of a survival probability estimated from Cox's regression model. *Biometrika* **73**, 722–724.
- Andersen, P.K., Christensen, E., Fauerholdt, L., and Schlichting, P. (1983), Measuring prognosis using the proportional hazards model. *Scand. J. Statist.* **10**, 49–52.
- Andersen, P.K. and Gill, R.D. (1982) Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100–1120.
- Cox, D.R. (1972) Regression models and life tables (with discussion). *J.R. Statist. Soc. B* **34**, 187–220.
- Cox, D.R. (1975) Partial likelihood. *Biometrika* **62**, 269–276.
- Efron, B. (1981) Censored data and the bootstrap. *J. Amer Statist. Assoc.* **76**, 312–319.
- Freireich, E.O. (1963) The effect of 6-mercaptopmine on the duration of steroid induced remission in acute leukemia. *Blood* **21**, 699–716.
- Gray, R.J. (1992) Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *J. Am. Stat. Assoc.* **87**, 942–951.
- Gustafson, P. (1998) Flexible Bayesian modelling for survival data. *Lifetime Data Analysis* **4**, 281–299.

- Hastie, T. and Tibshirani, R. (1993) Varying-coefficient models. *J. Roy. Stat. Soci. B* **55**, 757–796.
- Hinkley, D.V. (1988) Bootstrap methods. *J.R. Stat. Soc. B* **50**, 321–337.
- Kendall, M. (1975) *Rank Correlation Methods*. 4th ed., London and High Wycombe, Griffin.
- Kent, J.T. (1983) Information gain and a general measure of correlation. *Biometrika* **70**, 163–174.
- Kent, J.T. and O’Quigley, J. (1988) Measure of dependence for censored survival data. *Biometrika* **75**, 525–534.
- Korn, E.L. and Simon, R. (1990) Measures of explained variation for survival data. *Stat. Med.* **9**, 487–503.
- Maddala, G.S. (1983) *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge, U.K.
- Murray, S. and Tsiatis, A.A. (1996) Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics* **52**, 137–151.
- Murphy, S.A. and Sen, P.K. (1991) Time-dependent coefficients in a Cox-type regression model. *Stoch. Process. Applic.* **39**, 153–180.
- Oken, M.M., Leong, T., Lenhard, R.E., Greipp, P.R., Kay, N.E., Van Ness, B., Keimowitz, R.M., and Kyle, R.A. (1999) The addition of interferon or high dose cyclophosphamide to standard chemotherapy in the treatment of patients with multiple myeloma. *Cancer* **86**, 957–968.
- O’Quigley, J. and Flandre, P. (1994) Predictive capability of proportional hazards regression. *Proc. Natl. Acad. Sci. USA* **91**, 2310–2314.
- O’Quigley J., Flandre P., and Reiner E. (1999) Large sample theory for Schemper’s measures of explained variation in the Cox model. *J. Roy. Stat. Soc. D* **48**, 53–62.
- O’Quigley, J. and Xu, R. (2001) Explained variation in proportional hazards regression. *Handbook of Statistics in Clinical Oncology* J. Crowley, ed., 397–409. Marcel Dekker, Inc., New York.
- Rashid, S.A., O’Quigley, J., Axon, A.T.R., and Cooper, E.H. (1982) Plasma protein profiles and prognosis in gastric cancer. *Br. J. Cancer* **45**, 390–394.
- Sargent, D.J. (1997) A flexible approach to time-varying coefficients in the Cox regression setting. *Lifetime Data Analysis* **3**, 13–25.
- Schemper, M. (1990) The explained variation in proportional hazards regression. *Biometrika* **77**, 216–218.
- Schemper, M. (1994) Correction: “The explained variation in proportional hazards regression”. *Biometrika* **81**, 631.
- Schemper, M. and Kaider, A. (1997) A new approach to estimate correlation coefficients in the presence of censoring and proportional hazards. *Comp. Stat. and Data Anal.* **23**, 467–476.
- Schemper, M. and Stare, J. (1996) Explained variation in survival analysis. *Stat. Med.* **15**, 1999–2012.
- Schoenfeld, D.A. (1982) Partial residuals for the proportional hazards regression model. *Biometrika* **69**, 239–241.
- Sleeper, L.A. and Harrington, D.P. (1990) Regression splines in the Cox model with application to covariate effects in liver disease. *JASA* **85**, 941–949.
- Verweij, J.A. and Van Houwelingen, H.A. (1995) Time-dependent effects of fixed covariates in Cox regression. *Biometrics* **51**, 1550–1556.
- Xu, R. (1996) *Inference for the Proportional Hazards Model*. Ph.D thesis, University of California, San Diego.
- Xu, R. and Adak, S. (2001) Survival analysis with time-varying relative risks: a tree-based approach. *Methods of Information in Med.* **40**, 141–147.
- Xu, R. and O’Quigley, J. (1999) A R^2 type of measure of dependence for proportional hazards models. *Nonparametric Stat.* **12**, 83–107.
- Zucker, D.M. and Karr, A.F. (1990) Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *The Ann. Stat.* **18**, 329–353.

20 Constructing Prognostic Groups by Tree-Based Partitioning and Peeling Methods

*Michael LeBlanc, Erik Rasmussen, and
John Crowley*

CONTENTS

20.1	Introduction	366
20.2	Notation and Prognostic Models	367
20.2.1	Function Approximation	367
20.2.2	Extreme Region Summary	368
20.3	Tree-Based Methods	368
20.3.1	Constructing a Tree Rule	368
20.3.2	Splitting	371
20.3.3	Pruning and Selecting a Tree	372
20.3.3.1	Methods Based on Within-Node Deviance.....	372
20.3.3.2	Methods Based on Between-Node Separation	373
20.3.4	Further Recombination of Nodes.....	374
20.4	Data Peeling Methods	374
20.4.1	Regions of Interest	375
20.4.2	Region Refinement.....	376
20.4.3	Variable Selection and Peeling Fraction	377
20.4.4	Selecting Rules by Cross-Validation	378
20.4.5	Constructing Multiple Box Rules	380
20.5	Discussion	380
	Acknowledgments	381
	References	381

20.1 INTRODUCTION

Construction of prognostic subgroups of patients with often performed to help understand the association of patient characteristics with survival times and to aid in the design of clinical trials. Applications include the development of stratification schemes for future clinical trials and the identification of patients who are suitable for a cancer therapy targeted at a specific prognostic group.

The proportional hazards model of Cox¹ is a flexible tool for the study of covariate associations with survival time. It has been used to identify prognostic groups of patients by using the linear component of the model (prognostic index) or informally through counting up the number of poor prognostic factors corresponding to terms in the fitted model. However, the model does not directly lead to an easily interpretable description of patient prognostic groups. An alternative to using prognostic indices constructed from the proportional hazards model is a rule that can be expressed as simple logical combinations of covariate values. For example, an individual with some hypothetical type of cancer may have a poor prognosis if ($age \geq 60$) and ($serum\ creatinine \geq 2$) or ($serum\ calcium < 5$). This chapter presents two types of methodologies for constructing these logical rules for prognosis: 1) tree-based methods, which partition the data into multiple prognostic groups and 2) peeling methods, which sequentially refine the data into patient subsets with either very good or very poor prognosis.

Tree-based methods were formalized and extensively studied by Breiman, Friedman, Olshen, and Stone.² Trees have also recently been of interest in machine learning; one example is the C4.5 algorithm due to Quinlan.³ Tree-based methods recursively split the data into groups leading to a fitted model that is piecewise constant over regions of the covariate space. Each region is represented by a terminal node in a binary decision tree. Tree-based methods have been extended to censored survival data for the goal of finding groups of patients with differing prognoses.⁴⁻⁸ Further extensions and examples for tree-based rules in a wide variety of medical research problems were considered by Zhang and Singer.⁹ Some examples of tree-based methods for survival data in clinical studies are given in.¹⁰⁻¹²

Trees are useful methods for constructing multiple prognostic groups. However, in some cases the goal is to construct a simple decision rule to describe a single patient subset with either very good or very poor prognosis. For that objective, complementary techniques to tree-based methods have recently been developed. These methods are related to the patient rule induction method (PRIM) proposed by Friedman and Fisher.¹³ The basic idea of the PRIM algorithm and extensions is to describe a region in the covariate space corresponding to the most extreme expected outcome. PRIM was developed for uncensored data and works best with very large data sets but has been extended by LeBlanc, Jacobson, and Crowley¹⁴ to work better with clinical trials type data with low signal and also to survival data by LeBlanc, Moon, and Crowley.¹⁵ The main components of the algorithm are common between the published methods. Initially the entire data set is considered, and then a fraction α of the data is removed from either extreme of a variable distribution among all variables. This process, called peeling, is repeated until only a small fraction of the data remains. The data corresponding to a rule with sufficiently extreme

outcome are removed, and the remaining data are peeled again. The end result is a logical rule representing an extreme outcome group.

In this chapter we discuss the general methodological aspects of both tree-based partitioning and data peeling. We illustrate the methods with a large data set collected by the International Myeloma Foundation (IMF) for the goal of constructing a reliable staging system for multiple myeloma.¹⁶ Clinical and laboratory data were gathered on previously untreated myeloma patients from 17 institutions including sites in North America, Europe, and Asia. Patient characteristics were typical for symptomatic myeloma including age, sex, and clinical as well as laboratory parameters.

20.2 NOTATION AND PROGNOSTIC MODELS

We assume that X is the true survival time, C is the random censoring time, and Z is a p -vector of covariates. The observed variables are the triple $(T = X \wedge C, \Delta = I\{X \leq C\}, Z)$ where T is the time under observation and Δ is an indicator of failure. Given Z , we assume that X and C are independent. The data consist of a sample of independent observations $\{(t_i, \delta_i, z_i): i = 1, 2, \dots, N\}$ distributed as the vector (T, Δ, Z) . The survival probability at time t space

$$S(t|z) = P(X > t|z)$$

20.2.1 FUNCTION APPROXIMATION

Most survival data regression models vary smoothly as a function of the covariates. For instance, the proportional hazards model specifies the hazard function form as

$$\lambda(t|z) = \lambda_0(t) \exp(\eta(z))$$

where $\lambda_0(t)$ is a baseline hazard function and $\eta(z)$ is the logarithm of the relative risk. The relative risk function is sometimes referred to as the prognostic index and is typically a linear function of the covariates $\eta(z) = z'\beta$. More general additive models for the logarithm of the relative risk function are also used for modeling survival data. While additive function expansions are useful for variable interpretation, rules for patients with differing prognoses, which are typically of the form $\{z : \eta(z) \geq q\}$ or $\{z : \eta(z) \leq q\}$, are difficult to describe because they are a weighted combination of patient characteristics. A tree-based method is also a function approximation method, except that the piecewise constant approximating model is homogeneous over regions of the prediction space. A tree-based model for survival can be represented as

$$S(t|z) = \sum_{h=1}^H S_h(t) I\{z \in B_h\}$$

where B_h is a box shaped region in the predictor space, represented by a terminal node h , and the function $S_h(t)$ is the survival function corresponding to region B_h . H is the number of overlapping regions. Importantly, each terminal node can be described by a logical rule, for instance, $(z_1 < 3) \cap (z_2 \geq 6) \cap (z_5 < 2)$. With the

sample sizes typically available for clinical applications, a piecewise constant model can yield quite poor approximations to the conditional survival function, $S(t|z)$, which is likely a smooth function of the underlying covariates. Therefore, methods using smooth models such as linear Cox regression likely yield better approximations than piecewise constant tree-based methods. However, the primary motivation for using tree-based models is the potentially easy interpretation of decision rules describing multiple groups of patients, and such rules are not directly obtained by methods that assume $S(t|z)$ is a smooth function of the covariates.

20.2.2 EXTREME REGION SUMMARY

An alternative strategy does not attempt to model the entire regression function but instead describes the survival function in one region R ,

$$S(t|R) = S_R(t)I\{z \in R\}$$

corresponding to patients expected to have an extreme good or poor outcome. Unlike tree-based methods there is no attempt to describe the other regions. However, as with trees, R is defined by logic rules. An important attribute is the mass or fraction of the sample within the region R which can be denoted by

$$\beta(R) = \int_R dP_z$$

where dP_z represents the probability mass function associated with the covariates.

Estimation of the rules will be reviewed in this chapter. However, at the most basic level, tree-based regression constructs regions or boxes B_h by repeatedly partitioning the data, and the extreme region summaries are obtained via sequentially removing small fractions of data along coordinate axes. [Figure 20.1](#) and [Figure 20.2](#) show a schematic of each of the two procedures with two covariates.

The myeloma data from the International Myeloma Foundation presented in this chapter are from 2,677 patients with survival outcome and complete covariate data for *calcium* (CAL), *serum β_2 microglobulin* (B2M), *platelets* (PLT), *serum creatinine* (CREAT), *serum albumin* (ALB) and *percent plasma cells in the bone marrow* (BMPC).

20.3 TREE-BASED METHODS

20.3.1 CONSTRUCTING A TREE RULE

A tree-based model is developed by recursively partitioning the data. At the first step the covariate space is partitioned into two regions and the data are split into two groups. The splitting rule is applied recursively to each of the resulting regions until a large tree has been grown. Splits along a single covariate are used because they are easy to interpret. For an ordered covariate, splits are of the form $z_j < c$ or $z_j \geq c$, and for a nominal covariate splits are of the form $z_i \in S$ or $z_i \notin S$, where S is a nonempty subset of the set of labels for the nominal predictor z_j . Potential splits are evaluated for each of the covariates, and the covariate and split value resulting in the greatest reduction in impurity is chosen.

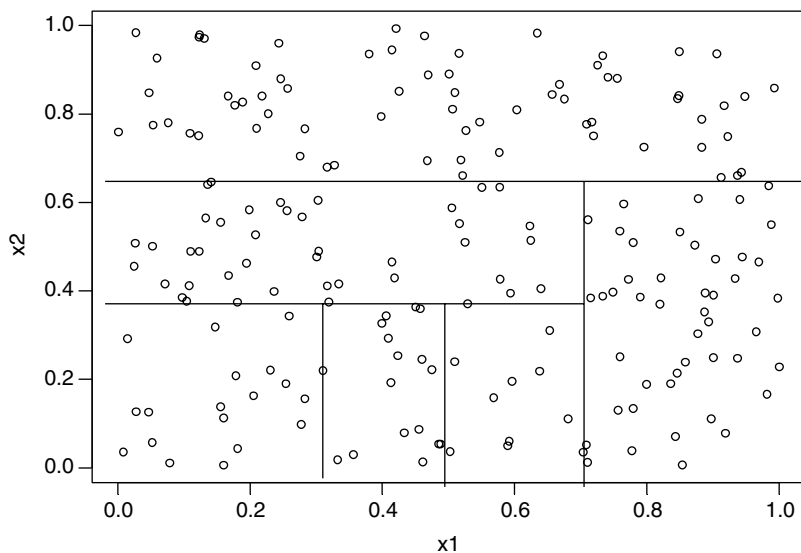


FIGURE 20.1 Schematic of partitioning of two-dimensional covariate space.

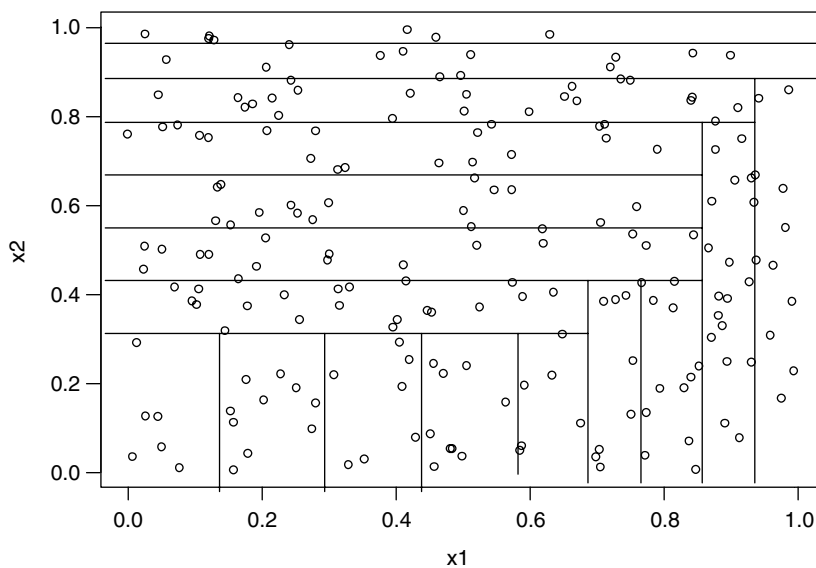


FIGURE 20.2 Schematic of peeling of two-dimensional covariate space.

The improvement for a split at node h into left and right daughter nodes $l(h)$ and $r(h)$ is

$$G(h) = R(h) - [R(l(h)) + R(r(h))]$$

where $R(h)$ is the residual error at node h . We assume $G(h) > 0$.

For uncensored continuous response problems, $R(h)$ is typically the mean residual sum of squares or mean absolute error. For survival data, it would be reasonable to use deviance corresponding to an assumed survival model. For instance, the exponential model deviance for node h is

$$R(h) = \sum_{i=1}^N 2 \left[\delta_i \log \left(\frac{\delta_i}{\hat{\lambda}_h t_i} \right) - (\delta_i - \hat{\lambda}_h t_i) \right]$$

where $\hat{\lambda}_h$ is the maximum likelihood estimate of the hazard rate in node h . Often an exponential assumption for survival times is not valid. However, a nonlinear transformation of the survival times may make the distribution of survival times closer to an exponential distribution. LeBlanc and Crowley⁷ investigate a full-likelihood method that is equivalent to transforming time by the marginal cumulative hazard function and using the exponential deviance.

However, most recursive partitioning schemes for censored survival data use the logrank test statistic of Mantel¹⁷ for $G(h)$ to measure the separation in survival times between two groups. Simulation studies of the performance of splitting with the logrank and other between-node statistics are given in Ref. 8 and 18.

Figure 20.3 shows the value of the logrank test statistic for groups defined by partitioning on serum β_2 microglobulin ($B2M$), ($B2M < c$), and ($B2M \geq c$) for range of cutpoints in the International Myeloma Foundation (IMF) data set. The largest logrank test statistic corresponds to a split at $c = 8.89$ and would lead to the first split in a tree-based model to be ($B2M < 8.89$) versus ($B2M \geq 8.89$).

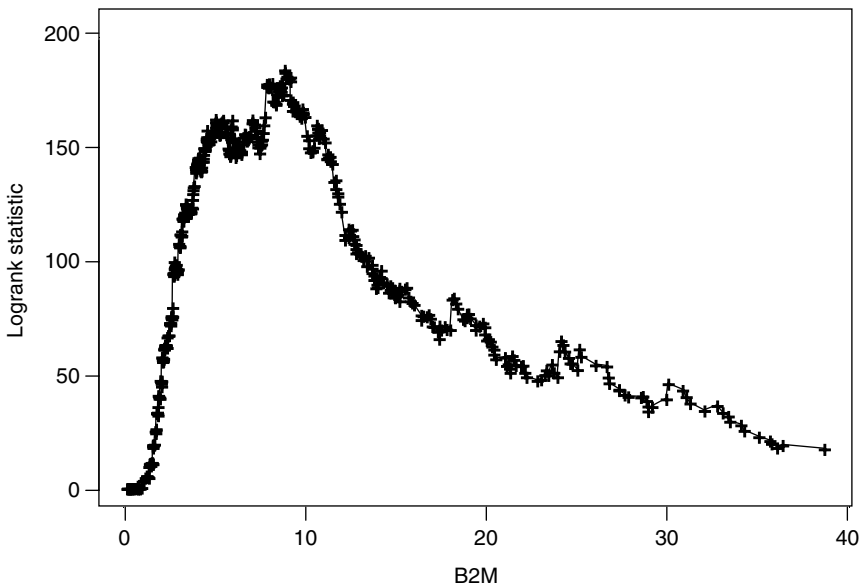


FIGURE 20.3 Logrank test statistics at observed values of serum β_2 microglobulin ($B2M$).

20.3.2 SPLITTING

If there are weak associations between the survival times and covariates, splitting on a continuous covariate tends to select splits that send almost all the observations to one side of the split. This is called end-cut preference by Breiman et al.² When growing survival trees, we restrict both the minimum total number of observations and the minimum number of uncensored observations within any potential node. This restriction is also important for prognostic stratification, because very small groups of patients are usually not of clinical interest.

It is also important that the splitting statistic can be calculated efficiently for all possible split points for continuous covariates. While the logrank test is relatively inexpensive to calculate, one way to improve computational efficiency is use a simple approximation to the logrank statistic which allows simple updating algorithms to consider all possible splits.¹⁹ Updating algorithms can also be constructed for exponential deviance; for instance see Ref. 19 and 20.

Figure 20.4 shows a tree grown on the IMF using the logrank test statistic for splitting, with a constraint of a minimum node size of 5% of the sample size. The tree has 15 terminal nodes. The logrank test statistic and permutation p -value are presented below each split in the tree. The p -value is calculated at each node by permuting the responses over the covariates and recalculating the best split at that node 1,000 times and then calculating the proportion of logrank test statistics greater than

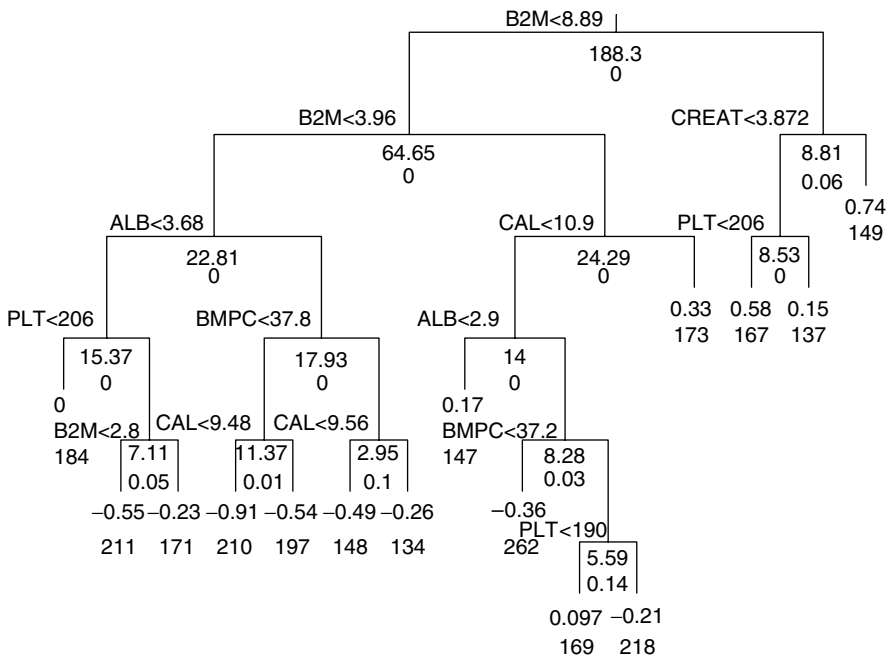


FIGURE 20.4 Unpruned survival tree. Below each split is the logrank test statistic and a permutation p -value. Below each terminal node is the logarithm of the hazard ratio relative to the leftmost node in the tree and the number of cases in the node.

the observe statistic. At each terminal node, the logarithm of the hazard ratio relative to the leftmost node and the number of cases falling into each terminal node are presented. The logarithm of the hazard ratio is obtained by fitting a Cox¹ model with dummy variables defined by terminal nodes in the tree. The worst prognostic group are patients with very high serum β_2 microglobulin ($B2M$) ($B2M \geq 8.89$) and high serum creatinine ($CREAT$) ($CREAT \geq 3.872$) and corresponds to an estimated logarithm of the hazard ratio relative to the prognostic group represented by the leftmost node equal to 0.74. While the minimum node size was set to be quite large (5% of the sample or approximately 134 observations), the logrank test statistics near the bottom of the tree (and permutation p -values) indicate there may be several nodes that should be combined to simplify the model.

20.3.3 PRUNING AND SELECTING A TREE

Two general methods have been proposed for pruning trees for survival data. The methods that use within-node error or deviance usually adopt the CART pruning algorithm directly.

20.3.3.1 Methods Based on Within-Node Deviance

In the CART algorithm, the performance of a tree is based on the cost complexity measure

$$R_\alpha(T) = \sum_{h \in \tilde{T}} R(h) + \alpha |\tilde{T}|$$

of the binary tree T , where $\tilde{T}$ is the set of terminal nodes, $|\tilde{T}|$ is the number terminal nodes, α is a nonnegative parameter, and $R(h)$ is the cost (often deviance) of node h .

A subtree (a tree obtained by removing branches) T_0 is an optimally pruned subtree for any penalty α of the tree T if

$$R_\alpha(T_0) = \min_{T' \leq T} R_\alpha(T')$$

where $\leq$ means is a subtree of, and T_0 is the smallest optimally pruned subtree if $T_0 \leq T''$ for every optimally pruned subtree T'' .

The cost complexity pruning algorithm obtains the optimally pruned subtree for any α . This algorithm finds the sequence of optimally pruned subtrees by repeatedly deleting branches of the tree for which the average reduction in impurity per split in the branch is small. The cost complexity pruning algorithm is necessary for finding optimal subtrees because the number of possible subtrees grows very rapidly as a function of tree size.

The deviance will always decrease for larger trees in the nested sequence based on the data used to construct the tree. Therefore, honest estimates of deviance for a new sample are required to select a tree that would have small expected deviance. If a test sample is available, the deviance for the test sample can be calculated for each of the pruned trees in the sequence using the node estimates from the training sample. For instance the deviance at a node would be

$$R(h) = \sum_{i \in N_h} 2 \left[\delta_i^T \log \left(\frac{\delta_i^T}{\hat{\lambda}_h t_i^T} \right) - (\delta_i^T - \hat{\lambda}_h t_i^T) \right]$$

where (t_i^T, δ_i^T) are the survival times and status indicators for test sample observations falling into node h , $z_i^T \in B_h$ for the tree and node estimate $\hat{\lambda}_h$ calculated from the learning sample.

However, a test sample is usually not available. Therefore, the selection of the best tree can be based on resampling-based estimates of prediction error (or expected deviance). The most popular method for tree-based models is the K-fold cross-validation estimate of deviance. The training data L are divided into K test samples L_k and training samples $L_{(k)} = L - L_k$, $k = 1, \dots, K$ of about equal size. Trees are grown with each of the training samples $L_{(k)}$. Each test sample L_k is used to estimate the deviance using the parameter estimates from the training sample $L_{(k)}$. The K-fold cross-validation estimate of deviance is the sum of the test sample estimates. The tree that minimizes the cross-validation estimate of deviance (or a slightly smaller tree) is selected. While K-fold cross-validation is a standard method for selecting tree size, it is subject to considerable variability; this is noted in survival data in simulations given in LeBlanc and Crowley.⁸ Therefore, methods such as those based on bootstrap resampling may be useful alternatives.²¹ One bootstrap method based on logrank splitting is given in the next section.

20.3.3.2 Methods Based on Between-Node Separation

LeBlanc and Crowley⁸ developed an optimal pruning algorithm analogous to the cost complexity pruning algorithm of CART for tree performance based on between-node separation. They define the split complexity of a tree as

$$G_\alpha(T) = G(T) - \alpha|S|$$

where $G(T)$ is the sum over the standardized splitting statistics $G(h)$ in the tree T :

$$G(T) = \sum_{h \in S} G(h)$$

where S represents the internal nodes T

A tree T_0 is an optimally pruned subtree of T for complexity parameter α if

$$G_\alpha(T_0) = \max_{T' \leq T} G_\alpha(T')$$

and it is the smallest optimally pruned subtree if $T_0 \leq T'$ for every optimally pruned subtree. The algorithm repeatedly prunes off branches with smallest average logrank test statistics in the branch. An alternative pruning method for trees based on the maximum value of the test statistic within any branch was proposed by Segal.⁶

Because the same data are used to select the split point and variable as used to calculate the test statistic, we use a bias-corrected version of the split complexity described above

$$G_\alpha(T) = \sum_{h \in S} G^*(h) - \alpha|S|$$

where the corrected split statistic is

$$G^*(h) = G(h) - \Delta^*(h)$$

and where the bias is denoted by

$$\Delta^*(h) = E_{Y^*}G(h; L^*, L) - E_{Y^*}G(h; L^*, L^*)$$

The function $G(h; L^*, L)$ denotes the test statistic where the data L^* were used to determine the split variable and value and the data L were used to evaluate the statistic. The function $G(h; L^*, L^*)$ denotes the statistic where the same data were used to pick the split variable and value and to calculate the test statistic. The difference $\Delta^*(h)$ is the optimism due to adaptive splitting of the data. We use the bootstrap to obtain an estimate $\hat{\Delta}^*(h)$ then we select trees that minimize the corrected goodness of split

$$\tilde{G}_\alpha(T) = \sum_{h \in S} (G(h) - \hat{\Delta}^*(h)) - \alpha |S|$$

$\tilde{G}_\alpha(T)$ is similar to the bias-corrected version of split complexity used in LeBlanc and Crowley⁸ except here we do the correction locally for each split conditional on splits higher in the tree. We typically choose a complexity parameter $\alpha = 4$. Note, if splits were not selected adaptively, an $\alpha = 4$. would correspond approximately to the 0.05 significance level for a split, and $\alpha = 2$ is in the spirit of the Akaike information criterion (AIC).²² Permutation sampling methods can also be used to add an approximate p -value to each split, conditional on the tree structure above the split to help the interpretation of individual splits.

Figure 20.5 shows a pruned tree based on the corrected goodness of split using 25 bootstrap samples with $\alpha = 4$. There are eight terminal nodes.

20.3.4 FURTHER RECOMBINATION OF NODES

Usually only a small number of prognostic groups are of interest. Therefore, further recombination of nodes with a similar prognosis from the pruned tree may be required. We select a measure of prognosis (for instance, hazard ratios relative to some node in the tree or median survival for each node) and rank each of the terminal nodes in the pruned tree based on the measure of prognosis selected. After ranking the nodes, there are several options for combining nodes in a pruned tree. One method is to grow another tree on the ranked nodes and only allow the second tree to select three or four nodes; another method is to divide the nodes based on quantiles of the data; and a third method is to evaluate all possible recombinations of nodes into V groups and choose the partition that yields the largest partial likelihood or largest V sample logrank test statistic. The result of recombining to yield the largest partial likelihood (third method) for a four-group combination of the pruned myeloma tree given in Figure 20.5 is presented in Figure 20.6.

20.4 DATA PEELING METHODS

Tree-based methods are effective for describing multiple prognostic strata. In this section we focus on a method for finding a single poor or good prognostic group with greater control of the relative outcome for patients in that group. In developing a clinical trial for a new aggressive therapy, one often must limit the study to only

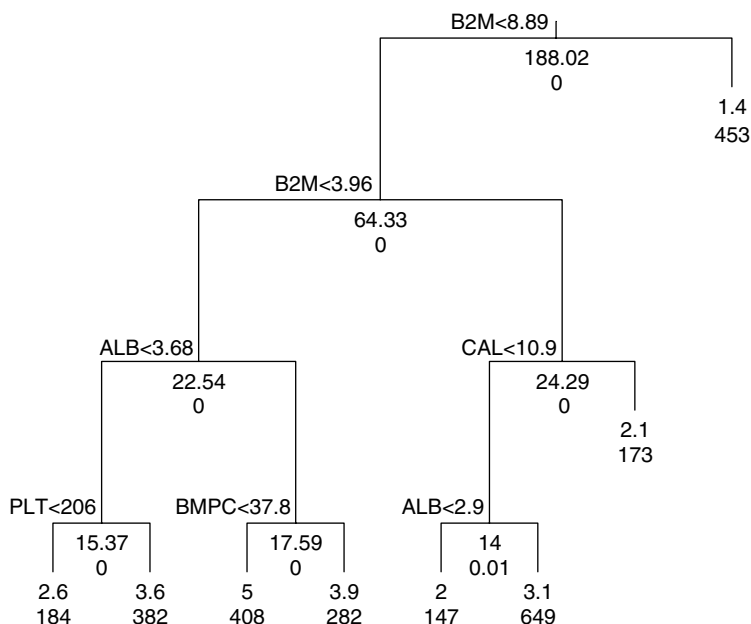


FIGURE 20.5 Pruned survival tree. Below each split is the logrank test statistic and permutation p -value. Below each terminal node is the median survival (in years), and underneath that is the number of patients represented by that node.

those patients with sufficiently poor prognosis appropriate for the toxicity associated with that therapy. Conversely, if a group of patients can be identified who have very good prognosis, one may want to investigate less toxic therapies for that group. However, the prognostic group must include a sufficient proportion of the patients with that disease to make patient accrual to the clinical trial feasible. Data peeling is a strategy for defining a prognostic group based on several covariates by repeatedly refining rules by removing data along the covariate axes. The method allows one to look at average patient outcome (median survival or k -year survival probability) as a function of the fraction of the sample represented by the rule.

20.4.1 REGIONS OF INTEREST

Interest focuses on a region R of the predictor space with extreme patient outcome values. Denote the functional of interest Q (e.g., mean, median, or 5-year survival probability) for that region by

$$Q(R) = Q_{z \in R}(z)$$

We assume a proportional hazards (PH) model with a conditional hazard function defined by

$$\lambda(t; z) = \lambda_0(t) \exp(\eta(z))$$

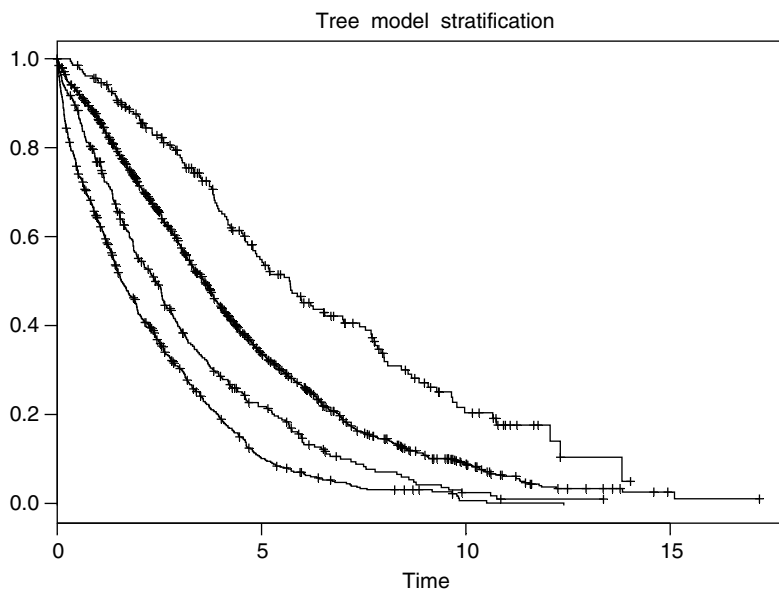


FIGURE 20.6 Survival for four prognostic groups derived from the pruned survival tree. The top curve corresponds to survival for node 3; the next curve corresponds to nodes 2, 4 and 6; the next to nodes 1 and 5; and the bottom curve to nodes 7 and 8 (numbered from the left on [Figure 20.5](#)).

where $\lambda_0(t)$ is an unknown baseline hazard function and $\eta(z) = \eta I\{z \in R\}$, with $I\{z \in R\}$ equal to 1 if the covariate value is in R or 0 otherwise.

To facilitate description, we focus on a region R that can be described by decision rules. For instance, we construct interpretable models based on boxes in the predictor space. If we assume the individual covariates, $z_1, \dots, z_p$ are ordered, then univariate rules are of the form $D_j = \{z : z_j > s_j\}$ or $D_j = \{z : z_j \leq s_j\}$ $j = 1, \dots, p$. Boxes can then be expressed as the intersection of the intervals,

$$B_k = \bigcap_{j=1}^p D_j$$

Quite general regions can be defined using unions of boxes in the predictor space. Such rules are sometimes referred to as rules of disjunctive normal form. A model for two groups can be represented by $\eta(z) = \eta I\{z \in R\}$ where $R = \cup_k B_k$ is a union over boxes.

20.4.2 REGION REFINEMENT

The peeling algorithm constructs rules (boxes) by repeatedly removing small amounts of data along the coordinate axes. Initially the entire data set is considered, and then a fraction α of the data is removed along the covariate axis, which leaves the largest parameter estimate for the remaining data in the box. The removal of data is continued until some minimum fraction of the observations remain in the box. The

fraction of data to be removed at each step is taken to be quite small so that the procedure can recover from bad local decisions. However, a minimum number of observations n_{min} must be removed at each step to control the variance of the procedure.

Since survival data tend to have low signal, we have found it advantageous to remove data only in one direction for each variable. We refer to this as directed peeling. The directions for refinement or data removal can be picked based on prior knowledge of the impact of the variables on disease outcome or based on signs of the coefficients from a fitted linear proportional hazards model or a semiparametric method such as local partial likelihood as developed by Gentleman and Crowley.²³

While k-year survival probability or median survival can be used to guide the peeling process, we have found it useful to use the hazard ratio from a Cox model since censoring may make those first quantities inestimable within some regions of the predictor space. The directed peeling algorithm changes the face of a box that maximizes the rate of increase in the estimated hazard ratio parameter associated with a box, $\eta(z) = \eta I\{z \in B\}$. Let

$$d_p = \frac{\hat{\eta}_{B_p} - \hat{\eta}_B}{\hat{\beta}(B) - \hat{\beta}(B_p)}$$

denote the change in the estimated regression parameter $\hat{\eta}_{B_p}$ from the current value $\hat{\eta}_B$ for a given proposed new box B_p obtained by changing a box face along one axis. A cutpoint is changed from s_p to $s_{p'}$ to modify the box $B_p = B \cap \{z_p \leq s_{p'}\}$ or $B_p = B \cap \{z_p > s_{p'}\}$ depending on the direction of the peeling. Because clinical data often include tied values for some covariates, one cannot remove an exact fraction of the data at each step. Therefore, we standardize the change in hazard ratio by the difference in support between the current and refined box $\hat{\beta}(B) - \hat{\beta}(B_p)$.

We note that the peeling process is computationally manageable. An upper bound on the number of steps for a peeling algorithm is $\log(\rho)/\log(1-\alpha)$ where ρ is the minimum fraction of the data of interest in a risk group.¹³ For example, for $\alpha = 0.1$ and $\rho = 0.05$, the maximum number of steps is 29. Focusing on the hazard ratio function allows a convenient way to select poor prognosis boxes as a function of the mass of the box. However, the median survival or proportion of sample in the prognostic group are useful summaries and are included as output from our peeling software.

As an example of the directed peeling process, we peel with respect to two variables *serum β_2 microglobulin (B2M)* and *serum albumin (ALB)* to identify a group of myeloma patients with particularly poor survival. The first 15 steps of the refinement sequence are given in Table 20.1 for rules of the form

$$B2M \geq c_1 \cap ALB < c_2$$

We call sequence of all the survival estimates the trajectory curve of the procedure.

20.4.3 VARIABLE SELECTION AND PEELING FRACTION

To limit the amount of data adaptation, we use variable selection to limit the number of variables used in peeling. Options are to select the top r variables based on the most significant univariate partial likelihood score tests or to limit peeling to variables that

TABLE 20.1

Sequence of Thresholds for the First 15 Steps of Peeling a Poor Prognostic Group for the Myelona Data Set ^a

Step	c_1 (<i>B2M</i>)	c_2 (<i>ALB</i>)	Median Survival
1	18.4	—Inf	3.17
2	12.4	—Inf	3.25
3	9.8	—Inf	3.39
4	8.3	—Inf	3.42
5	8.3	2.5	3.53
6	8.3	2.8	3.55
7	8.3	2.9	3.62
8	7.4	2.9	3.70
9	6.6	2.9	3.73
10	6.0	2.9	3.78
11	5.5	2.9	3.82
12	5.2	2.9	3.85
13	4.8	2.9	3.85
14	4.6	2.9	3.92
15	4.6	3.1	3.95

^a The Rules Are of the Form $sb2m \geq c_1 \cap alb < c_2$.

have been selected by multivariable linear proportional hazards modeling. For instance, one can use a forward step-wise model building strategy to select the variables to be used in the subsequent peeling. This simple preselection strategy also constrains the complexity and yields more easily interpreted rules.

Once the variables have been selected, the only tuning parameters that need to be adjusted are the fraction and minimum number of observations to remove at each step of the peeling process and the minimum fraction of the data to be considered in a smallest final group. As a default, we chose to remove 10% ($\alpha = 0.1$) of the current number of observations available for peeling (minimum 10 observations) at each step and set a smallest final group size to 5% ($\rho = 0.05$) of the original data set.

20.4.4 SELECTING RULES BY CROSS-VALIDATION

Simulations have shown that while selecting a small number of covariates for peeling dramatically reduces estimation bias due to selection,¹⁴ it is still useful to have less biased estimates for rule selection. Given that one rarely has a test sample available to select the rule, we propose selecting rules using K-fold cross-validation. Bootstrap bias correction methods would be an alternative as we described for tree-based methods.

We use K-fold cross-validation as previously described for tree-based regression. Peeling is applied to each of the training samples $L_{(k)}$; each test sample L_k is used to estimate the quantity of interest Q_k from the model grown on the training sample $L_{(k)}$.

For each training sample $L_{(k)}$, the trajectory curve $Q_k(\cdot)$ is defined as a piecewise constant curve $Q_k(x) = Q_k(\hat{\beta}_{l(k)})$ for $\hat{\beta}_{l(k)+1} \leq x < \hat{\beta}_{l(k)}$ where $\hat{\beta}_{l(k)}$ is the support of the l -th box in the k -th training sample trajectory. We define $\hat{\beta}_0 = 1$, $Q_k(1)$ as the

functional calculated on the entire k -th test sample and $\hat{\beta}_{l(k)+1} = 0$, where $L(k)$ is the smallest box in the k -th trajectory.

The cross-validation estimate is the average over the individual curves indexed as a function of the fraction of the sample

$$Q^{cv}(\hat{\beta}_j) = \frac{1}{K} \sum_{k=1}^K Q_k(\hat{\beta}_j)$$

where $\hat{\beta}_j$ is the observed support from the entire training sample for box B_j and $Q_k(\cdot)$ is functional of the k -th test sample applied to the trajectory from the k -th training sample. It is our experience that K-fold cross-validation can yield quite variable results. Therefore, for the peeling method we repeat K-fold cross-validation several times and average. The peeling trajectory for the myeloma data set and the 10-fold cross-validated estimates (averaged five times) are presented in Figure 20.7. In this case, the cross-validation estimates suggest limited selection bias. Here we suppose we are interested in patients with at least 4.5 years median survival. If we apply the peeling rule and cross-validation, the resulting rule is

$$\text{RULE 1: } B2M \leq 3.5 \cap ALB > 3.71$$

which corresponds to a cross-validated estimate of median survival equal to 4.5 years and represents approximately 20% of the patients.

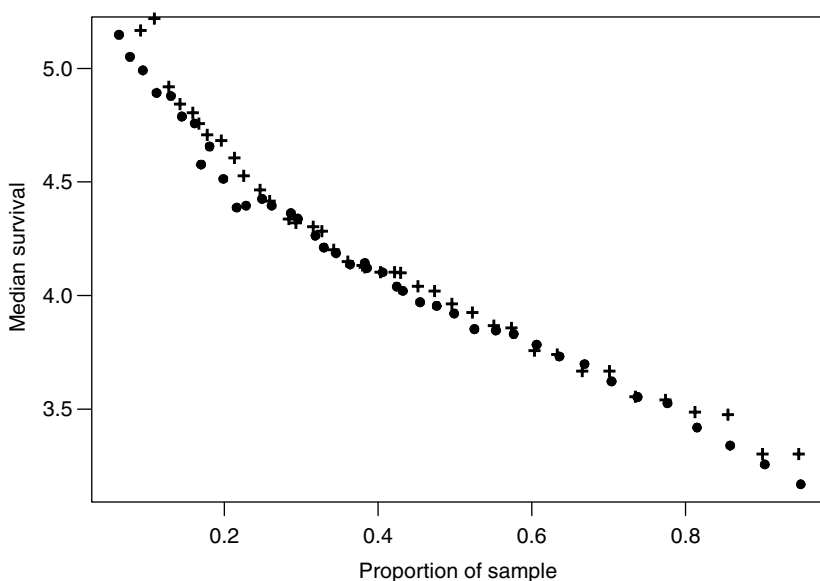


FIGURE 20.7 First peeling trajectory on the myeloma data focusing on a good prognosis group (• represents training sample estimates and + represents averaged 10-fold cross-validation estimates).

20.4.5 CONSTRUCTING MULTIPLE BOX RULES

We have described the construction of a single box trajectory. The peeling process can be applied again to find additional patients with sufficiently extreme (poor or good) outcome. Therefore, the rules are refined through the addition (union) of other boxes. One removes the data corresponding to the first rule before the next rule is constructed. At the M th iteration the box is constructed by using data not contained in any of the previous boxes

$$\left\{ (t_i, \delta_i, z_i) \mid z_i \notin \bigcup_{m=1}^{M-1} B_m \right\}$$

The estimated median survival from the Kaplan–Meier estimator (denoted by KM_{MED}) for a box at iteration M is

$$Q_M = KM_{MED} \left\{ t_i, \delta_i \mid z_i \in B_M \cap z_i \notin \bigcup_{m=1}^{M-1} B_m \right\}$$

and the support in the box excluding previous boxes is

$$\beta_M = \text{ave} \left\{ I \{ z_i \in B_M \cap z_i \notin \bigcup_{m=1}^{M-1} B_m \} \right\}$$

where *ave* denotes average. We applied this strategy to the myeloma data to determine if additional patients could be identified with a median survival of greater than 4.5 years using the ten-fold cross-validation estimates (averaged over five replications). While no additional group of patients could be identified with a survival greater than 4.5 years, the following rule depending on four variables (*serum β_2 microglobulin (B2M)*, *platelets (PLT)*, *serum creatinine (CREAT)* and *serum albumin (ALB)*) describes the next 10% of patients with best overall survival,

$$\text{RULE 2: } PLT > 208 \cap CREAT \leq 1.43 \cap B2M \leq 3.5 \cap ALB > 3.0$$

Combined with patients identified with RULE 1, the good prognostic group represents 30% of myeloma patients with estimated median survival of 4.2 years. One can also identify good prognosis patients from the tree-based model. In that case, 15% patients were identified with a survival better than 4.5 years, but then adding in the patients with the next best survival leads to a group with 4.4 year median survival consisting of 26% of patients. This somewhat smaller fraction of patients compared to peeling is mostly due to the discreteness of the tree-based rules, which partition the data in large fractions.

20.5 DISCUSSION

Both the tree-based and peeling methods describe prognostic groups using decision rules that we believe are useful for interpretation and in clinical settings. The primary objective of tree-based methods with censored survival data is to provide an easy to understand description of multiple prognostic groups of patients. By contrast, the peeling method focuses on a single extreme outcome group. For problems

where one wants to control either the proportion of patients in a poor/good prognostic group or the prognosis (e.g., median survival or 5-year survival probability), peeling can be a useful complement to tree-based methods, which tend to give more discrete answers.

Software implementing tree-based methods based on logrank splitting using SPLUS is available from the first author. Other software has been written for tree-based modeling for survival data. The RPART program implements a recursive partitioning based on within-node error that includes an exponential model-based survival tree.²⁴ Another implementation of software based on logrank test statistics based on Segal⁶ is available from that author and is called TSSA. Software implementing peeling with survival data is also available from the first author of this chapter.

Tree-based partitioning and peeling are not the only methods available for constructing prognostic decision rules. Two other related methods have recently been developed for constructing logic prognostic rules. Logic regression is a very general technique for constructing logic or Boolean rules for regression models using binary covariates.²⁵ The algorithm uses stochastic searches to avoid getting caught in local optima. Another technique, Extreme regression, like peeling, is suitable for describing subsets of patients with either good or poor prognosis when the covariates are largely continuous. Unlike peeling, this new method models the entire regression function but akin to peeling leads to simple logic rules to describe extreme groups.²⁶

ACKNOWLEDGMENTS

This work was supported by the US NIH through R01-CA90998 and R01-CA074841. The authors want to thank Mark Blitzer for help in preparation of this manuscript.

REFERENCES

1. Cox, D.R. (1972) Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.
2. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984), *Classification and Regression Trees*. Wadsworth, Belmont, CA.
3. Quinlan, J.R. (1993) *C4.5 Programs for Machine Learning*, Morgan Kaufman San Francisco, CA.
4. Gordon, L. and Olshen, R.A. (1985) Tree-structured survival analysis, *Cancer Treatment Reports*, **69**, 1065–1069.
5. Ciampi, A., Hogg, S., McKinney, S., and Thiffaut, J. (1988) RECPAM: a computer program for recursive partition and amalgamation for censored survival data, *Computer Methods and Programs in Biomedicine*, **26**, 239–256.
6. Segal, M.R. (1988) Regression trees for censored data, *Biometrics*, **44**, 35–48.
7. LeBlanc, M. and Crowley, J. (1992) Relative risk regression trees for censored survival data *Biometrics*, **48**, 411–425.
8. LeBlanc, M. and Crowley, J. (1993) Survival trees by goodness of split, *Journal of the American Statistical Association*, **88**, 457–467.

9. Zhang, H. and Singer, B. (1999) *Recursive Partitioning in the Health Sciences*, Springer-Verlag, New York.
10. Albain, K., Crowley, J., LeBlanc, M., and Livingston, R. (1990) Determinants of improved outcome in small cell lung cancer: an analysis of the 2580 patient Southwest Oncology Group data base, *Journal of Clinical Oncology*, **8**, 1563–1574.
11. Ciampi, A., Thiffault, J., Nakache, J.-P., and Asselain, B. (1986). Stratification by stepwise regression, correspondence analysis, and recursive partition, *Computational Statistics and Data Analysis*, **4**, 185–204.
12. Kwak, L.W., Halpern, J., Olshen, R.A., and Horning, S.J. (1990) Prognostic significance of actual dose intensity in diffuse large-cell lymphoma: results of a tree-structured survival analysis, *Journal of Clinical Oncology*, **8**, 963–977.
13. Friedman, J. and Fisher, N. (1999) Bump hunting in high dimensional data (with discussion), *Statistics and Computing*, **9**, 123–162.
14. LeBlanc, M., Jacobson, J., and Crowley, J. (2002) Partitioning and peeling for constructing patient outcome groups, *Statistical Methods for Medical Research*, **11**, 247–274.
15. LeBlanc, M., Moon, J., and Crowley, J. (2005) Adaptive risk group refinement, *Biometrics*, **61**, 370–378.
16. Greipp, P., San Miguel, J., Durie, B., Avet-Loiseau, H., Facon, T., Fonseca, R., Rasmussen, E., Cavo, M., Crowley, J., Attal, M., Barlogie, B., Beksac, M., Bladé, J., Boccadoro, M., Child, J., Harousseau, J., Kyle, R., Lahuerta, J., Ludwig, H., Morgan, G., Moreau, P., Powles, R., Shimizu, K., Shustik, C., Sonneveld, P., Tosi, P., Turesson, I., and Westin, J. (2005) An international staging system (ISS) for multiple myeloma (MM), *Journal of Clinical Oncology*, **23**, 3412–3420.
17. Mantel, N. (1966) Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemotherapy Reports*, **50**, 163–170.
18. Crowley, J., LeBlanc, M., Gentleman, R., and Salmon, S. (1995) Exploratory methods in survival analysis. In *Analysis of Censored Data*, IMS Lecture Notes — Monograph Series 27, H.L. Koul and J.V. Deshpande, eds., 55–77, Institute of Mathematical Statistics, Hayward, CA.
19. LeBlanc, M. and Crowley, J. (1995) Step-function covariate effects in the proportional-hazards model, *The Canadian Journal of Statistics*, **23**, 109–129.
20. Davis, R. and Anderson J. (1989) Exponential survival trees, *Statistics in Medicine*, **8**, 947–962.
21. Efron, B. and Tibshirani R. (1997) Improvements on cross-validation: the .632 bootstrap, *Journal of the American Statistical Association*, **92**, 548–560.
22. Akaike, H. (1974) A new look at model identification, *IEEE Transactions on Automatic Control*, **19**, 716–723.
23. Gentleman, R. and Crowley, J. (1991) Local full likelihood estimation for the proportional hazards model, *Biometrics* **47**, 1283–1296.
24. Therneau, T.M. and Atkinson, E.J. (1997) An introduction to recursive partitioning using the RPART routines. Technical report, Mayo Foundation (distributed in PostScript with the RPART package).
25. Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2003) Logic regression. *Journal of Graphical and Computational Statistics*, **12**, 475–511.
26. LeBlanc, M., Moon, J., and Kooperberg, C. (2005) Extreme regression, *Biostatistics*, In Press.

21 Clinical Monitoring Based on Joint Models for Longitudinal Biomarkers and Event Times

Donna Pauler Ankerst and Dianne M. Finkelstein

CONTENTS

21.1	Introduction	383
21.2	Prostate Cancer	384
21.3	Joint Model	384
21.4	Application	387
21.5	Final Remarks	390
Appendix		391
	Prior Distribution for PSA Model	391
	MCMC Sampler for PSA Model	392
	Sensitivity and Goodness-of-Fit for PSA Model	393
References		393

21.1 INTRODUCTION

Biomarkers that are both sensitive and specific for tumor regrowth or disease progression are increasingly becoming available and routinely monitored during the course of regular follow-up of patients treated for cancer. Some specific examples are prostate-specific antigen (PSA) for prostate cancer, the CA19-9 antigen for pancreatic cancer, and the CA125 antigen for ovarian cancer. Discovery of new biomarkers through use of modern high-dimensional proteomic techniques is rapidly gaining speed (Early Detection Research Network, 2004). Obtained by a simple blood test, biomarkers are a less expensive and invasive alternative to other clinical monitoring modalities such as ultrasound or bone scan. Retrospective studies for several cancers and affiliated biomarkers have found that increases or changes in the biomarker can precede clinical evidence of recurrence of disease by time periods ranging from 1 month to 2 years depending on the context.¹⁻³ Following some therapies, biomarkers initially decline, and the rates of such initial declines provide early independent prognostic indicators of

clinical outcome; for examples in PSA following chemotherapy for advanced hormone-refractory prostate cancer see Ref. 4. Optimized predictions of clinical event times then should incorporate prognostic information from the full posttherapy biomarker sequence. Such methods should also utilize the repeated measures to adjust for measurement error inherent in a biomarker process. Joint statistical models for longitudinal and event time data provide promising tools that satisfy these criteria.

Joint models have been extensively used for a variety of objectives, including accounting for nonignorable missing data and assessment of the correlation between biomarkers and disease onset or recurrence in the presence of measurement error. For examples, see.^{5–11} In this chapter we focus on their use for obtaining predictions of time until recurrence or mortality for censored individuals or individuals still in the process of active follow-up. To illustrate the methods we consider a specific application in the monitoring of PSA levels following radiotherapy to monitor for prostate cancer recurrence. In Section 21.2. we provide details of the application. In Section 21.3 we outline a joint model, Markov chain Monte Carlo (MCMC) fitting algorithm, and an extension for computing predictions for censored individuals. In Section 21.4. we show results of the analysis applied to postradiotherapy serial PSA levels from 1011 patients treated for prostate cancer at Massachusetts General Hospital (MGH). Finally, in Section 21.5. we make concluding remarks.

21.2 PROSTATE CANCER

For monitoring patients treated by radiotherapy for first incidence of prostate cancer, focus has centered on PSA. Following radiotherapy, PSA levels typically decline slowly over a period that can extend up to two years and either remain at a plateau if the patient remains cancer-free or elevate if the cancer recurs. [Figure 21.1](#) contains typical PSA series from four patients from Massachusetts General Hospital (MGH), two that were observed to recur (top) and two that were not (bottom).

For our application to investigate the prognostic significance of PSA for the prediction of prostate cancer recurrence, we consider all postradiotherapy PSA levels from 1011 patients treated for prostate cancer at MGH during the years 1988 to 1995, and followed through 1999. Follow-up time for the group of MGH patients was on average 5 years, with a range of 3 days to 11 years. Patients were monitored at irregular intervals during the follow-up period, with levels of PSA recorded at each clinical visit. Biopsies and ultrasound examinations were performed approximately annually, and whenever there was an abnormal PSA value, above 4 ng/mL, or symptomatic indications of cancer recurrence. Clinical recurrence was defined based on detection from a bone scan, endorectal MRI, cystoscopy, or other diagnostic exam. There were 258 observed clinical recurrence times, with 74.5% of the follow-up times censored, and a total of 5332 PSA measurements.

21.3 JOINT MODEL

The form of the statistical joint model we use is a shared parameter selection model, which has also been employed in.^{7–10} Denote by $y_i(t)$ the biomarker measurement at

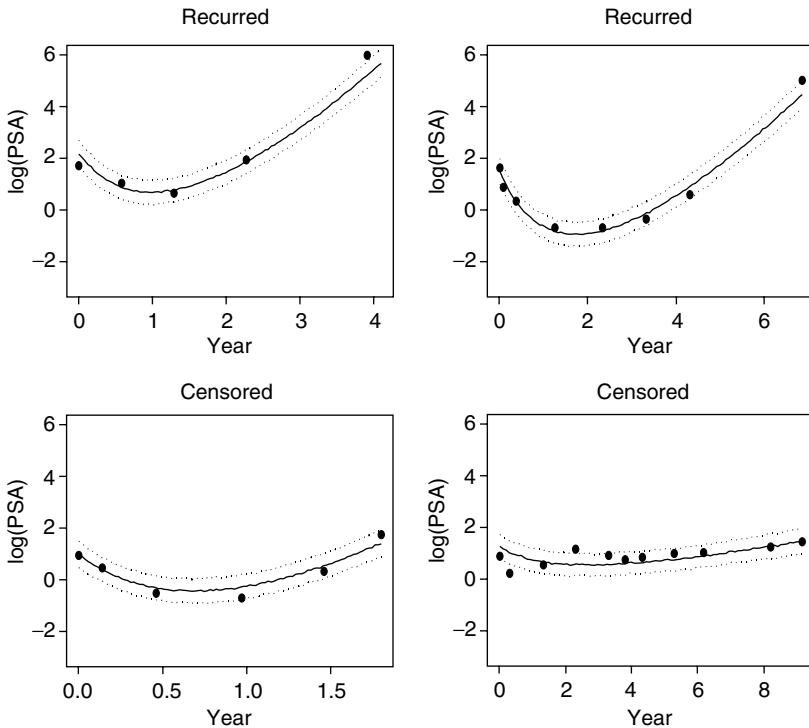


FIGURE 21.1 Posterior predictive mean curves with pointwise single standard error curves for two patients who recurred (top) and two who were censored (bottom) during follow-up. The recurrence times for the patients on top are indicated on the graph by vertical lines.

time t , s_i the time of clinical event or censoring, and d_i the indicator equal to one if the event is observed and zero otherwise, for patient i , for $i = 1, \dots, I$. Some transformation such as the logarithm is typically applied so that the distribution of the data are better approximated by a normal distribution. The data for each patient are the vector (s_i, d_i, y_i) , where $y_i = (y_i(t_{i1}), \dots, y_i(t_{ini}))$ is the vector of biomarker measurements for the i th patient. Let θ_i denote a vector of subject-specific random effects and θ a vector of population parameters. We specify a random coefficient selection model for (s_i, d_i, y_i) as

$$p(y_i, s_i, d_i, \theta_i, \theta) = p(y_i | \theta_i, \theta) p(s_i, d_i | \theta_i, \theta) p(\theta_i | \theta) p(\theta) \quad (21.1)$$

where observations corresponding to different patients i are independent. To complete the longitudinal specification, we assume a normal distribution for $y_i(t)$ with mean function $\mu_i(t | \theta_i)$ and error variance σ^2 : $y_i(t) \sim N(\mu_i(t | \theta_i), \sigma^2)$. At the second stage of the longitudinal model component we assume $\theta_i \sim N_q(\mu, D)$ where q is the dimension of θ_i and $N_q(\mu, D)$ denotes the q -variate normal distribution with mean vector μ and

variance–covariance matrix D . For the event time data (s_i, d_i) we assume a Cox proportional hazards model, which specifies that the hazard of failing at time t is equal to

$$\lambda_0(t) \exp[\eta'x(\theta_i, t)] \quad (21.2)$$

where $\lambda_0(t)$ denotes the baseline hazard function, η a vector of regression coefficients, and $x(\cdot)$ a vector function of time and patient-specific parameters, to be determined later. A Bayesian approach requires a full likelihood and hence a flexible parametric model for $\lambda_0(t)$. Therefore, we break the interval $[0, T]$, where T is a time point beyond the maximum observed or censored time, into K subintervals $(t_k, t_{k+1}]$, for $k = 1, \dots, K$, and assume a step function with heights λ_k over each interval to obtain $\lambda_0(t) = \sum_k \lambda_k I_{(t_k, t_{k+1}]}(t)$, where $I_{(a, b]}(x)$ is the indicator function that equals 1 if $x \in (a, b]$ and 0 otherwise. Assumption (21.2) with a piecewise linear baseline hazard function yields a piecewise exponential model for the recurrence time distribution.^{12,13} To complete the Bayesian specification of the model, prior distributions for the population parameters θ are defined, which may vary from non-informative to informative across the components of θ , as required to obtain a proper posterior distribution. Determination of the prior distributions requires an understanding of the identifiability of each component based on the likelihood. In any case sensitivity analyses to the choice of prior distributions should accompany the analysis. As an example of prior specification, the [appendix](#) to this chapter contains details of the prior distributions for the MGH PSA example outlined in Section 21.4.

Model selection specific to the application at hand comes into play in the choice of mean trajectory $\mu_i(t|\theta_i)$ for the longitudinal measurements and choice of covariates $x(\theta_i, t)$ for the survival component (21.2). Other adaptations to the model are possible, such as replacing the normal distribution at the first stage of the longitudinal model component with a t distribution, to better accommodate outliers, and should be considered following goodness of fit assessments. The choice of the form $\mu_i(t|\theta_i)$ depends on the nature of the biomarker trajectory for the specific treatment under consideration. We give a specific example of a nonlinear function that has been motivated for the MGH PSA example in Section 21.4. Other applications have used simple linear trends⁹ and single change-point functions.¹⁴ Dependence of the random effects θ_i on baseline covariates z such as age, stage of disease, or treatment can be incorporated into the model by replacing individual components of μ with linear terms β'_z . Similarly the choice of covariates $x(\theta_i, t)$ for inclusion in the Cox hazard rate (21.2) can take many forms. A common choice is the current underlying biomarker value, represented by the time-varying function $\mu_i(t|\theta_i)$. Another option is the current velocity or rate of change, expressed by $\mu'_i(t|\theta_i)$ where f' denotes the derivative of f . Still other options include these quantities defined at baseline, $t = 0$. Given the wealth of possibilities, it is helpful to use an external model selection tool to select the optimal choice of covariates for prediction of the survival data. One option is to fit only the longitudinal component of the model to the longitudinal biomarker measurements, save the individual posterior means of the random effects, construct all possible covariates for consideration for each individual using the point estimates, and run a survival model selection program using a forward, backward, or

stepwise search. The covariates selected in the optimized model can then be used for the joint model (21.1). An example of the approach is given in Section 21.4.

The Markov chain Monte Carlo (MCMC) method for drawing samples from the posterior distribution comprises a series of Gibbs and Metropolis–Hastings steps.¹⁵ The [appendix](#) gives the MCMC scheme for the MGH PSA application, which we have found to work well and which generalizes quite easily to other models.

For censored individuals or those currently under active follow-up it is possible to obtain an estimate of the posterior predictive distribution of the time to a clinical event by embedding a single extra step into the MCMC scheme. Specifically one can obtain a draw t_i from $p(t_i|\theta, \theta, \{d_i = 0\}, \{t_i > s_i\}, \bar{\mathcal{D}})$, where $\bar{\mathcal{D}} = (\bar{y}, \bar{s}, \bar{d})$ denotes observed data from all patients, at each iteration of the sampler by use of the inversion method. Let $F_c(t_i)$ denote the cumulative distribution for the full conditional distribution of t_i given current values of θ_i and θ , the data $\bar{\mathcal{D}}$ and conditional on $\{d_i = 0\}$ and $\{t_i > s_i\}$. Under our model assumptions, F_c can be explicitly written. The inversion method is implemented as follows. First draw $u \sim U[0,1]$ and then find t_i satisfying $F_c(t_i) = u$. The latter step is accomplished using a one-dimensional root-finding numerical subroutine.

21.4 APPLICATION

For the MGH PSA application, we take the logarithm of PSA measurements as the longitudinal observation $y_i(t)$. The following mean model, recommended by,¹⁶ tracks the prostradiotherapy log PSA trajectory fairly well:

$$\mu_i(t|\theta) = \alpha_i + \beta_i t - \gamma_i \log(t + 1) \quad (21.3)$$

In (21.3) α_i denotes the log PSA value at end of radiotherapy; positive γ_i controls the rate of decline of PSA immediately post-therapy; and positive β_i controls the rate of rise of PSA for large t . Examples of the types of curves fitted by (21.3) are shown in [Figure 21.1](#). At the second stage of the longitudinal model component we specify a population distribution for the patient-specific parameters $(\alpha_i, \beta_i, \gamma_i)$ as $(\alpha_i, \log \beta_i, \log \gamma_i)' \sim N_3(\mu, D)$. The random-effects distribution constrains β_i and γ_i to be nonnegative for each individual, thereby imposing a mean trajectory that is either flat or convex over time and allowing identification of the time of PSA nadir, a quantity of interest to clinical oncologists.¹⁷ From (21.3) and the constraints $\beta_i > 0, \gamma_i > 0$, the time of nadir for each patient is given by $t_{\text{nadir},i} = \gamma_i/\beta_i - 1$, and the PSA value at nadir is given by $\mu_i(t_{\text{nadir},i}|\alpha_i, \beta_i, \gamma_i)$. We set prior distributions for the population parameters $\theta = (\sigma^2, \mu, D, \eta)$ as defined in the [appendix](#).

We perform a model selection procedure for the covariates $x(\cdot)$ in (21.2) by first fitting the longitudinal model only and then using posterior medians of individual random effects to form the possible covariates. In addition to model-based covariates we also include age and stage of disease. We then use SAS Proc PHREG¹⁸ with the stepwise option to find the model with lowest value of Schwarz's Bayesian information criterion:¹⁹ $\text{SBC} = -2 \times [\text{maximized log partial likelihood} - \text{no. of parameters} \times \log \text{sample size}]$, where sample size denotes the number of observed events,

TABLE 21.1
Regression Coefficients, Hazard Ratios, and *p*-Values for the Cox Proportional Hazards Model Using Posterior Medians of Longitudinal Parameters

Coefficient	Predictor	Estimate	Hazard ratio	p-value
SBC=2473.6				
η_1	β_i	0.67	1.95	<0.0001
η_2	γ_i	−0.26	0.77	<0.0001
η_3	$\mu_i(t)$	0.84	2.32	<0.0001
η_4	$\mu'_i(t)$	−0.45	0.64	0.04
η_5	$I(t_i > t_{\text{nadir},i}) \times \mu'_i(t)$	1.34	3.82	<0.0001
η_6	$\beta_i \times \mu_i(t)$	−0.13	0.88	<0.0001

^a $\mu_i(t)$ denotes $\mu_i(t|\alpha_i, \beta_i, \gamma_i)$

258. The top model chosen, shown in Table 21.1, contains four main effects and two interactions.

The MCMC sampler used to fit the MGH PSA joint model is outlined in the [appendix](#). We ran the MCMC sampler for a 5000 iteration burn-in period and an additional 5000 iterations, saving every tenth observation for a sample output of 500 elements. To assess convergence we informally examined traceplots of the MCMC sampler for all population parameters, including the λ_k parameters used to define the baseline hazard. The plots indicated convergence had been obtained. Assessments of goodness of fit and sensitivity analyses are outlined in the appendix. Fitted curves are overlaid on the observed data for the four patients in [Figure 21.1](#).

Posterior estimates for all population parameters, excluding the λ_k s, are shown in [Tables 21.2](#) and [21.3](#). The posterior estimates of the survival population parameters are not as tightly estimated as the longitudinal parameters, as seen by their increased length of posterior intervals. The posterior medians of η_1 , η_2 , η_3 , and η_6 are of the same sign and of similar magnitude as the maximum likelihood estimates in Table 21.1. The parameters corresponding to the dependence of recurrence on the derivative of the mean function, η_4 and η_5 , now become “insignificant” in the sense that their posterior intervals include 0.

[Figure 21.2](#) shows posterior predictive densities for the time to failure for the two censored patients at the bottom of [Figure 21.1](#), respectively. The patient on the left is followed for just under 2 years and has a PSA trajectory that is slightly rising toward the end of this period. The posterior predictive density for time to failure has a mean at 2.19 years, which is 4.7 months after the current censoring time of 1.8 years. The probability that this patient will fail in the next 6 and 12 months following censoring is 0.68 and 0.96, respectively, which is quite high. The evidence indicates that this patient will recur soon and should be closely observed. In contrast, the patient on the right has a much longer follow-up period of 9.17 years, which is cancer free. There is a very slight increasing trend in PSA over the last five years of the follow-up period. Accordingly, the posterior predictive distribution of the recurrence time is quite diffuse over the next 3 years, where all distributions were truncated at 12 years, and the probability that this patient will fail in the next 6 and 12 months is only 0.20 and 0.35, respectively. The evidence indicates that this patient is cured.

TABLE 21.2

Posterior Medians, 95% Posterior Intervals Defined by the 2.5% and 97.5% Posterior Quantiles, and Length of the Posterior Intervals from the MCMC Sampler for the Longitudinal Parameters

Parameter	Median	(2.5%, 97.5%)	Length
μ_1	1.51	(1.44, 1.58)	0.14
μ_2	0.10	(0.01, 0.19)	0.18
μ_3	1.23	(1.17, 1.28)	0.11
σ^2	0.17	(0.16, 0.18)	0.02
d_{11}	1.22	(1.09, 1.35)	0.26
d_{22}	1.42	(1.27, 1.60)	0.33
d_{33}	0.62	(0.55, 0.69)	0.14
d_{12}	0.77	(0.66, 0.88)	0.22
d_{13}	0.56	(0.48, 0.65)	0.17
d_{23}	0.92	(0.81, 1.03)	0.22

$^a\mu_1, \mu_2, \mu_3$ denote the population means of α_i , $\log \beta_i$ and $\log \gamma_i$, respectively; d_{11} denotes the population variance of α_i , d_{12} denotes the population covariance of α_i and $\log \beta_i$ and so forth.

TABLE 21.3

Posterior Medians (Hazard Ratios), 95% Posterior Intervals Defined by the 2.5% and 97.5% Posterior Quantiles, and Length of the Posterior Intervals from the MCMC Sampler for the Longitudinal Parameters in the Cox Component of the Model

Parameter	Predictor	Median (hazard ratio)	(2.5%, 97.5%)	Length
η_1	β_i	0.78(2.18)	(0.45, 1.12)	0.67
η_2	γ_i	-0.17(0.84)	(-0.43, -0.02)	0.41
η_3	$\mu_i(t)$	0.89(2.44)	(0.67, 1.11)	0.44
η_4	$\mu'_i(t)$	0.36(1.43)	(-0.47, 1.93)	2.40
η_5	$I(t_i > t_{\text{nadir},i}) \times \mu'_i(t)$	0.70(2.01)	(-0.99, 1.95)	2.94
η_6	$\beta_i \times \mu_i(t)$	-0.14(0.87)	(-0.18, -0.11)	0.07

$^a\mu_i(t)$ denotes $\mu_i(t|\alpha_i, \beta_i, \gamma_i)$

We randomly split the 1011 MGH patients into a subset of 758 patients, called the training set, and 253 patients called the test set. In the 253 patients comprising the test set, there were 75 observed events, termed cases, and 178 censored observations, termed controls. For all patients in the test set we set the censoring indicators equal to zero and calculated the probability that they would fail in the next 12 months based on their serial PSA history by analyzing their censored data with the full data from the 758 patients of the training set. We compared posterior estimates of the probability of recurring in the next 12 months between the 75 cases and 178 controls of the test set. We would expect the 75 cases to have higher probabilities of recurring in the next 12 months. For a given patient i with censored data and draws

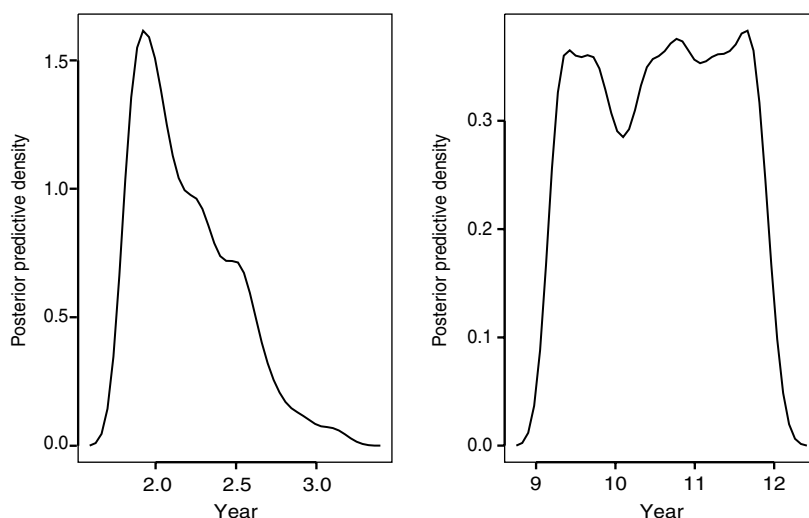


FIGURE 21.2 Posterior predictive densities of the time to recurrence for the two censored patients at the bottom of Figure 21.1.

$t_i^1, \dots, t_i^M$ from the posterior distribution of the time to recurrence, the posterior estimate of the probability of failing in the next 12 months is given by $1/M \sum_{m=1}^M I(t_i^m < 12 \text{ months})$. Figure 21.3 shows histograms for the censored (left) and recurred (right) patients comprising the test set. The median 12-month failure probability was 0.09 (95% confidence interval (0.014, 0.61)) for the censored individuals and 0.34 (0.015, 0.99) for the recurred cases. We note that there are a substantial number of low probabilities for recurred cases, which diminishes the sensitivity of the probability rule when used as a diagnostic test. We investigated the PSA trajectories and model fits for the 19 recurrences of the test set with probabilities of recurrence within 12 months less than 0.2. Ten of the 19 patients had very quick recurrence times after radiotherapy and hence only one or two PSA measurements available. The model-based predictions are more accurate for patients with a larger number of PSA measures.

21.5 FINAL REMARKS

Currently calculations of event time predictions are computationally intensive. C code for the MCMC algorithms used in this chapter is available upon request, but it is our hope that Web-based calculators for predicting risk based on entry of current biomarker histories will soon be available, as has been the case for the Gail model (<http://brca.nci.nih.gov/brc/questions.htm>), BRCAPRO (<http://astor.som.jhmi.edu/brcapro/>) and prostate cancer nomograms (<http://www.mskcc.org/mskcc/html/10088.cfm>).

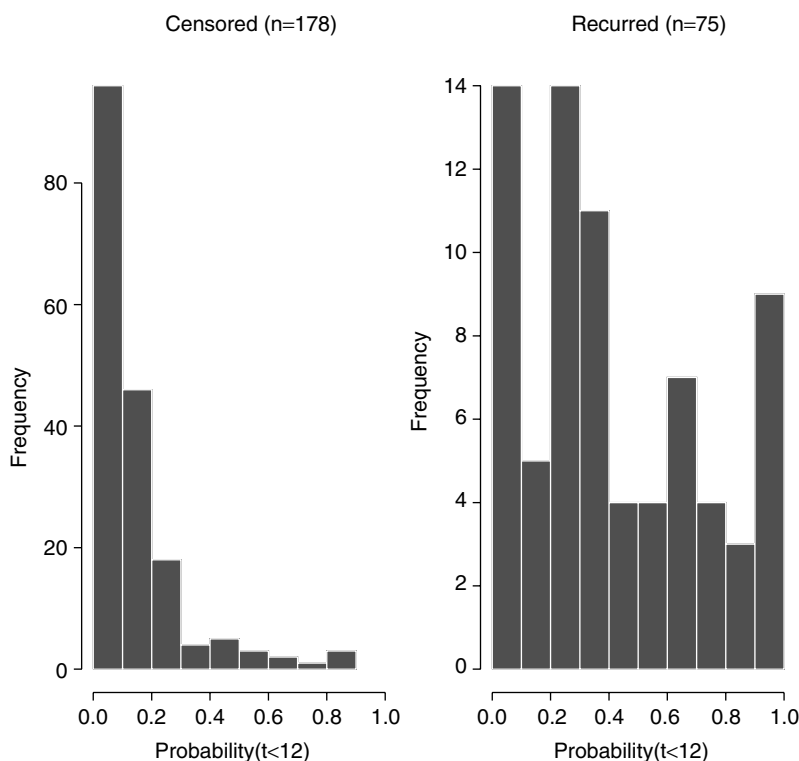


FIGURE 21.3 Histograms of posterior probability of failure within 12 months for censored (left) and recurred (right) cases in the 253 patient test set.

APPENDIX

PRIOR DISTRIBUTION FOR PSA MODEL

For σ^2 we specify the improper prior, $p(\sigma^2) \propto 1/\sigma^2$, corresponding to σ locally uniform.²⁰ For μ we choose a trivariate normal prior with mean vector $(0, -1, -1)$ and variance–covariance matrix equal to the identity matrix I_3 , and for D , an inverse Wishart prior with location matrix $S = 0.1I_3$ and degrees of freedom $\rho = 13$. The hyperparameters were chosen by comparing prior predictive longitudinal PSA curves to the observed PSA trajectories to ensure that prior curves are more dispersed than observed ones. Since the total number of observed recurrences is sufficiently large for moderate dimensions of η , we set a flat improper prior for η . Finally, we specify independent gamma $(0.1, 0.1)$ priors with mean 1 and standard deviation 10 for the heights λ_k , for $k = 1, \dots, K$.

MCMC SAMPLER FOR PSA MODEL

The MCMC scheme iterates between sampling from the full conditional distribution of all patient-specific and population parameters. The full conditional distribution of μ given all other parameters in the model is multivariate normal; that of D given all other parameters is inverse Wishart; and the conditional distributions of elements of $\{\lambda_k, k = 1, \dots, K\}$ are independent gamma distributions. Samples from the full conditional distributions of all other parameters are not conjugate, and so are obtained by the method of Metropolis–Hastings.¹⁵

The basic Metropolis–Hastings algorithm is as follows. For sampling a parameter X from a distribution $p(X)$, a proposal density $q(X)$ is sought that closely mimics $p(X)$ but for which it is easy to sample from directly. If X is the current value of the parameter during a MCMC iteration, a candidate value X' is drawn from $q(X)$. The candidate value X' is then accepted with probability

$$\alpha = \min \left\{ 1, \frac{p(X')q(X)}{p(X)q(X')} \right\} \quad (21.4)$$

If X' is not accepted, the current value X is retained.

The full conditional distribution for the vector of Cox regression parameters, η , is proportional to the product of the likelihood component for η conditional on the fixed current values of all other parameters in the model and the prior distribution of η . Because the prior distribution for η is proportional to a constant, its full conditional distribution is proportional to the likelihood component:

$$\prod_i \exp[d_i \eta' x_i(s_i, \alpha_i, \beta_i, \gamma_i)] \exp[-\sum_k I(s_i > t_k) \lambda_k \int_{t_k}^{\min(t_{k+1}, s_i)} \exp[\eta' x_i(t, \alpha_i, \beta_i, \gamma_i)] dt] \quad (21.5)$$

For a proposal density $q(\cdot)$, we use a multivariate normal density with mean and variance covariance matrix determined by the preliminary analysis used to determine the covariate matrix $x(\cdot)$. We use the maximum likelihood estimate (m.l.e.) of η for the mean and the asymptotic variance–covariance matrix evaluated at the m.l.e. for the variance-covariance matrix of the proposal density. This choice of proposal density for the MCMC sampler yields an approximate acceptance rate of 20%.

To sample from the full conditional distributions of α_i , β_i , and γ_i we also use the Metropolis–Hastings algorithm.¹⁵ For α_i we use as proposal density a normal distribution that is the product of the contribution to α_i arising from the likelihood component for individual i and the conditional normal prior distribution. For β_i and γ_i we use as proposal densities the contribution to the likelihood for the i th individual, which is normal, or the conditional normal prior distribution if there are insufficient data from the individual to form an identifiable likelihood contribution for the parameter. In the MCMC sampler, these proposal densities yield high acceptance rates, in the range of 80 to 90%, for individuals with three or more longitudinal PSA measurements.

SENSITIVITY AND GOODNESS-OF-FIT FOR PSA MODEL

We assessed sensitivity of the estimates in [Tables 21.2](#) and [21.3](#) to the choice of prior for the λ_k s comprising the baseline hazard. We refit the model using gamma (0.1, 0.1) priors, which have mean 1 and variance 100, and obtained posterior medians for all population parameters within 0.02 of the estimates in the tables. We checked the appropriateness of the longitudinal mean model by inspecting posterior predictive plots of individual mean curves, as shown in [Figure 21.1](#). For all of the patients checked, the curves mapped the observed trajectories closely, with most of the datapoints falling within the single standard error bands. We checked for dependency of the random effects α_i , β_i , and γ_i on stage of disease and baseline age by scatterplots of posterior medians of these effects versus stage and age. We found no trends with age and a slight increasing linear trend with increasing stage for all three random effects. Similar trends across all three random effects are to be expected due to posterior correlations between these variables. The dependency was slight with multiple R^2 s from the individual regressions approximately 0.08. The trend was driven primarily by the 19 patients with stage 4 disease. Because of the slight evidence we decided not to include stage as a regressor in the population mean μ for the random effects distribution. We checked the proportional hazards assumption for the survival portion by examining standardized Schoenfeld residuals, as prescribed in.²¹ Posterior residual plots indicated that the assumptions of the proportional hazards model approximately held.

REFERENCES

1. Rustin, G. J., Nelstrop, A. E., Tuxen, MK, and Lambert, H. E. Defining progression of ovarian carcinoma during follow-up according to CA125: a North Thames Ovary Group study. *Ann. of Oncol.* 1996;7:361–364.
2. American Society for Therapeutic Radiology and Oncology. Consensus statement: guidelines for PSA following radiation therapy. *Int. Radiation Oncol. Bio. Phys.* 1997;37:1035–1041.
3. Safi, F., Schlosser, W., Falkenreck, S., and Beger, H.G. Prognostic value of CA 19-9 serum course in pancreatic cancer. *Hepatogastroenterology* 1998;45:253–259.
4. Bubley, G.J., Carducci, M., Dahut, W., Dawson, N., Daliani, D., Eisenberger, M., Figg, W.D., Freidlin, B., Halabi, S., Hudes, G., Hussain, M., Kaplan, R., Myers, C., Oh, W., Petrylak, D.P., Reed, E., Roth, B., Sartor, O., Scher, H., Simons, J., and Sinibaldi, V., Small EJ, Smith MR, Trump DL, Vollmer R, Wilding G. Eligibility and response guidelines for phase II clinical trials in androgen-independent prostate cancer: recommendations from the prostate-specific antigen working group. *J Clin. Oncol* 1999;17:3461–3467.
5. Pawitan, Y. and Self, S. Modeling disease marker processes in AIDS. *J. Ame. Stati. Assoc.* 1993;88:719–726.
6. Little, R.J.A. Modeling the drop-out mechanism in repeated-measures studies. *J. Ame. Stat. Associ.* 1995;90:1112–1121.
7. Faucett, C.L. and Thomas, D.C. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Stat. Med.* 1996; 15:1663–1685.
8. Henderson, R., Diggle, P., and Dobson, A. Joint modelling of longitudinal measurements and event time data. *Biostatistics* 2000;1:465–480.

9. Tsiatis, A.A., and Davidian, M.A. Semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* 2001; 88:447–458.
10. Wang, Y., and Taylor J.M.G. Jointly modelling longitudinal and event time data, with application to acquired immunodeficiency syndrome. *J. Am. Stat. Assoc.* 2001; 96:895–905.
11. Law, N.G., Taylor, J.M.G., and Sandler, H. The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics* 2002;3:547–563.
12. Friedman M. Piecewise exponential models for survival data with covariates. *Ann. Stat.* 1982;10:101–113.
13. Aitkin, M., Anderson, D., Francis, B., and Hinde, J. *Statistical Modeling in GLIM*. Oxford: Oxford University Press, 1989.
14. Pauler, D.K., Finkelstein, D.M. Predicting time to prostate cancer recurrence based on joint models for non-linear longitudinal biomarkers and event time outcomes. *Stat. Med.* 2002;21:3897–3911.
15. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. Equations of state calculations by fast computing machine. *J. Chem. Phys.* 1953; 21:1087–1091.
16. Lu, J. and Tercilla, O. Estimation of growth curve models for partially unclassified data: an application to PSA screening after radiotherapy. In *ASA Proc. Biometrics Section* 1996;388–393.
17. Kuban, D.A., El-Mahdi, A.M., and Schellhammer, P.F., PSA for outcome prediction and posttreatment evaluation following radiation for prostate cancer: do we know how to use it? *Seminars in Rad. Onco.* 1998;8:72–80.
18. SAS Institute, Inc. *SAS/STAT User's Guide*, Version 8. Cary, NC:SAS Institute Inc., 1999.
19. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* 1978;6:461–464.
20. Jeffreys, H. *Theory of Probability*, 3rd ed. Oxford: Clarendon Press, 1961.
21. Grambsch, P.M., Therneau, T.M. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994;81:515–526.

Part V

High-Throughput Data and Bioinformatics

22 Some Practical Considerations for Analysis of Spotted Microarray Data

Li Hsu, James R. Faulkner, Douglas Grove, and Donna Pauler Ankerst

CONTENTS

22.1	Introduction	397
22.1.1	Description	398
22.1.2	Normalization.....	398
22.1.3	Analysis for Matched Samples	399
22.1.4	Multiple Testing	399
22.1.5	Datasets	400
22.2	Data Preprocessing or Normalization	400
22.2.1	Background Intensity Subtraction	401
22.2.2	Location Adjustment	403
22.2.3	Within- and Interarray Scale Adjustment	405
22.3	Analysis	409
22.3.1	ANOVA for Ratios	409
22.3.2	ANOVA for Intensities.....	410
22.3.3	Comparison of ANOVA Methods	411
22.4	Multiple Comparisons	412
22.5	Final Remarks	416
	Acknowledgment.....	416
	References	416

22.1 INTRODUCTION

Spotted microarrays are now commonly used for analysis of gene expression levels and DNA copy numbers for a variety of clinical and laboratory investigations. The ability to simultaneously measure the several thousands of genetic transcript levels or changes in

DNA copy number affords the opportunity to characterize complex biological processes at a level of detail not previously feasible; see, for example,¹⁻⁵

22.1.1 DESCRIPTION

A description of the cDNA microarray provides an example of the technological details. First cDNA is spotted on the microarrays using a print-tip, with typically 16 or 32 print-tips per array. Next mRNA from a sample of interest and a reference sample are reverse-transcribed into cDNA and are labeled differentially with two fluorescent dyes, for example Cy3 (green) and Cy5 (red). Two pools of cDNA are then mixed and washed over the array. The array is scanned for Cy3 and Cy5 fluorescent intensities, which reflect the relative amount of mRNA transcript for a given gene. The data consist of Cy3 and Cy5 measurements for every spot on every array, with each spot corresponding to transcripts, in the form of cDNA, encoded by a distinct gene.

Currently there are two variations to the usual cDNA microarrays described above. The first one spots cDNA onto the microarray but uses DNA rather than cDNA for the samples during the hybridization. So instead of measuring the transcript levels with cDNA, the array actually measures the relative copy number changes in the sample of interest compared with the normal sample.⁵ Advantages of this method include the widespread availability of cDNA clone-sets and the possibility of performing parallel analysis of changes in DNA copy number and expression levels using the same set of genes. However, there are also limitations in the use of cDNA clones for detecting copy number changes in genomic DNA. It enables only the detection of aberrations in known genes. Because it is the cDNA spotted onto the microarray, the hybridization of DNA to the cDNA array may not be optimal. For example, it may not be sensitive enough to detect low-copy number gains or losses.⁶ Another variation of the spotted microarray prints bacterial artificial clones (BAC), which are short DNA sequences, on the array^{4,7} and then hybridizes DNA samples onto the array. Individual spots on this array correspond to specific BAC clones rather than genes. This type of array is often called array-based comparative genomic hybridization or array-CGH. It overcomes some of the shortcomings of cDNA arrays, such as possible inadequate hybridization of DNA samples to the cDNA array, thus allowing for better detection and mapping of localized changes, such as gains or losses. A special feature of array-CGH data is possible high spatial correlations of changes among adjacent BACs.

22.1.2 NORMALIZATION

The technical processes underlying spotted arrays are subject to many sources of systematic and random errors, including differential incorporation and emission properties of the recording device, print-tip variation and wear, and uneven hybridization of the cDNA.⁸⁻¹¹ Unless properly adjusted for, such sources of variation will confound the primary analyses of interest. The preprocessing of the raw data to remove systematic variation is often called normalization and can be achieved by ensuring that the location and spread of intensity levels across dyes and arrays are comparable. There are a number of approaches for normalizing the data,

for example nonparametric modeling of ratios as a function of spot intensity levels⁸ or identification of a subset of conserved genes or BACs for use in normalizing the rest of the spots on the arrays.^{12,13} Not surprisingly, different normalization methods might yield different analysis results. The first objective of this chapter is to evaluate various normalization methods and assess their impact on the subsequent data analyses.

22.1.3 ANALYSIS FOR MATCHED SAMPLES

The data that arise from spotted microarrays consist of paired measurements at thousands of spots. In the commonly used reference design, one member of the pair corresponds to the sample of interest and the other to a reference sample. For this design it is natural to treat each spot as a matched set and to use the ratio to quantify the relative expression levels of the two samples. This partially adjusts for any spot effect in the analysis as the quantity of DNA that was printed to the spot is common for both samples. Alternatively, one may analyze the matched data in an unmatched fashion, analyzing the absolute level of intensity for each channel individually.¹⁴ In this case, the spot effect needs to be accounted for explicitly in the analysis to avoid confounding effects due to the matched samples. Gains in efficiency of the latter approach because of smoothing parameterizations can be offset by biases if the modeling assumptions are not correct.¹⁵ The second objective of this chapter is to study whether there is a difference between ratio-based and absolute intensity level-based data analyses of spotted microarray data.

22.1.4 MULTIPLE TESTING

With a single microarray experiment yielding intensity levels from thousands of genes or clones, an immediate concern is multiple testing. Bonferroni correction is simple and perhaps the most well-known method for accounting for multiple comparisons. It achieves strong control of the family-wise error rate (FWER), that is, control of the probability of at least one error in the family for any combination of true and false hypotheses. The approach can be highly conservative in the presence of strong dependence among genes or clones. Other less conservative procedures have been developed for providing strong control of the FWER, such as the Westfall–Young step-down procedure.¹⁶ Recently a number of researchers^{17,18} have suggested that rather than controlling for the FWER, controlling the proportion of genes falsely declared differentially expressed or BACs with aberrations might be more suitable in array studies. The proportion of falsely rejected genes or clones among the total rejected ones is called the false discovery rate (FDR). Given the same rejection region in the high-dimensional test statistic, the FDR yields an error rate less than or equal to that of the FWER under the null hypotheses for all genes or clones. In other words, procedures that control the FDR might have better power for rejecting the null hypotheses than those that control the FWER. The positive FDR (pFDR) is another variation of the FDR, where one estimates the FDR conditional upon positive findings having occurred. Undoubtedly, new measures will be increasingly used in this fast-growing field. However, the validity of these procedures under general dependence remains

under active investigation. Ge et al.¹⁹ provided a comprehensive review and comparison of these multiple comparison procedures in the context of gene expression data analysis. But little is known about how these procedures perform for the array-CGH data, where spatial correlation along chromosomes is an issue. To complement the Ge et al.¹⁹ study, our third objective is to compare the performance of various multiple comparison procedures on array-CGH data from breast tumors.

22.1.5 DATASETS

We illustrate the methods underlying these objectives using three spotted array datasets. The first one is a mouse experiment aimed to examine the normal variation of expression levels in kidney among genetically identical mice using cDNA arrays.²⁰ In this experiment, the total RNA was extracted from the kidneys of six male mice. A reference sample consisting of combined RNA from all samples including RNA from two other organs, the liver and testis, was cohybridized with the RNA from each kidney sample. Four separate microarrays with dye-swap were performed on each mouse's sample, resulting in a total of 24 arrays (6 mouse $\times$ 4 replicates). This dataset will be used to illustrate the normalization methods as well as the comparison between ratio-based and intensity-based methods for analysis.

The second and third datasets use array-CGH with no replicates. The second dataset is an X-titration experiment in which cell lines with varying numbers of X chromosomes 1, 3, 4, and 5 were each cohybridized with a normal female genomic DNA, resulting in 4 arrays. The expected ratios for the clones are 1/2, 3/2, 4/2, and 5/2, corresponding to the number of X-chromosomes, whereas non-X chromosome BAC clones are expected to have a ratio of 1. We will use this dataset as a benchmark dataset to examine the performance of various data preprocessing steps because we know the true underlying expected ratio for each clone. The third dataset is used to compare gene copy number changes between two subtypes of ductal breast tumors, estrogen receptor (ER) positive versus ER negative.⁷ ER negative tumors are often associated with poor prognosis and poor response to hormonal therapy.⁷ DNA samples from 14 ER positive and 13 ER negative flow-sorted tumor cells were assayed with a normal female DNA sample and hybridized on the array-CGH. This dataset will be used to compare the performance of various multiple comparison procedures.

Given the objectives outlined above, the remainder of the chapter proceeds as follows. Section 22.2 describes and compares methods for normalization; Section 22.3 compares the intensity versus ratio-based approaches for data analysis; and Section 22.4 compares multiple comparison procedures. Practical guidelines are given in each of the corresponding sections.

22.2 DATA PREPROCESSING OR NORMALIZATION

A hybridized spotted microarray is scanned to produce Cy3 and Cy5 fluorescence intensity measurements on a large collection of pixels on the array. These fluorescence intensities correspond to the levels of hybridization of the two samples to the cDNA or DNA sequences spotted onto the array. These intensities are usually stored as 16-bit images. A number of microarray image analysis packages have become

available. Yang et al.²¹ compared various image packages and found that the choice of segmentation procedure seemed to have little effect on results, whereas the background adjustment could substantially increase the intra- and inter-array variability for low intensity spots. Most image software²⁰ provides quality assessments and flags for potentially defective spots. GenePix imposes a criterion that the spot has to have at least 6 foreground pixels greater than 4 standard deviations above background as a way to ensure that the foreground intensity level is at least as large as the background intensity. It then discards flagged spots, which to some degree removes spots of very low intensity. GenePix Pro²² is used for the examples in this chapter.

After extraction of fluorescence intensities questions at issue are: (1) Should one subtract local background intensities from the foreground? (2) Which location-adjustment should be used? (3) Should one adjust for the varying scales of intensities within and across arrays? We will address each of these questions in the following subsections after introducing some notation.

We denote R for red and G for green for the intensity levels of the experimental and reference channels, respectively, and take the logarithmic transformation base 2 of the ratios R/G . Base 2 is chosen because intensities are typically integers between 0 and $2^{16} - 1$ and base 2 yields strictly positive values between 0 and 16. Our experiences with the array data indicate that the logarithm adequately transforms the data to more accurately follow a normal distribution. A common plot that we will use for examining the distribution of intensities for both channels is the M/A plot, where M is the log-intensity ratio, $M = \log_2 R/G$ and A is the average of log-intensities, $A = (\log_2 R + \log_2 G)/2$. The M/A plot is a 45° rotation of a scatter-plot of $\log_2 R$ versus $\log_2 G$, which allows for better visual inspection. For ease of presentation we drop the term $\log_2$. Finally, we focus exclusively on the normalization of ratios. Normalized absolute intensity levels for each channel can be obtained from the normalized ratios M and overall intensities A of each spot by a back transformation:¹⁰

$$\log_2 R = A + \frac{M}{2} \quad \text{and} \quad \log_2 G = A - \frac{M}{2} \quad (22.1)$$

22.2.1 BACKGROUND INTENSITY SUBTRACTION

The motivation for background adjustment is the belief that the measured foreground intensity for a spot includes not only the binding of the target to the probe but also a nonspecific hybridization or other chemicals on the glass. To obtain the intensity for the spot that is specific to the target, a commonly used measure is to subtract the local background intensity surrounding the spot from the foreground. This procedure relies critically on the assumption that the intensity of the regions surrounding the spot actually measures the nonspecific binding or reflects the chemical property on the glass for that spot. It is not clear to us, however, how this assumption could ever be verified. As an example, we empirically examined the performance of ratios with and without background intensity subtraction using our X-titration array-CGH data. Figure 22.1(a) shows M/A plots of X-titration arrays with and without background subtraction. As one can see, background subtraction tends to give quite noisy estimates for low-intensity spots. We also calculated the standard deviations and t -statistics of ratios of the

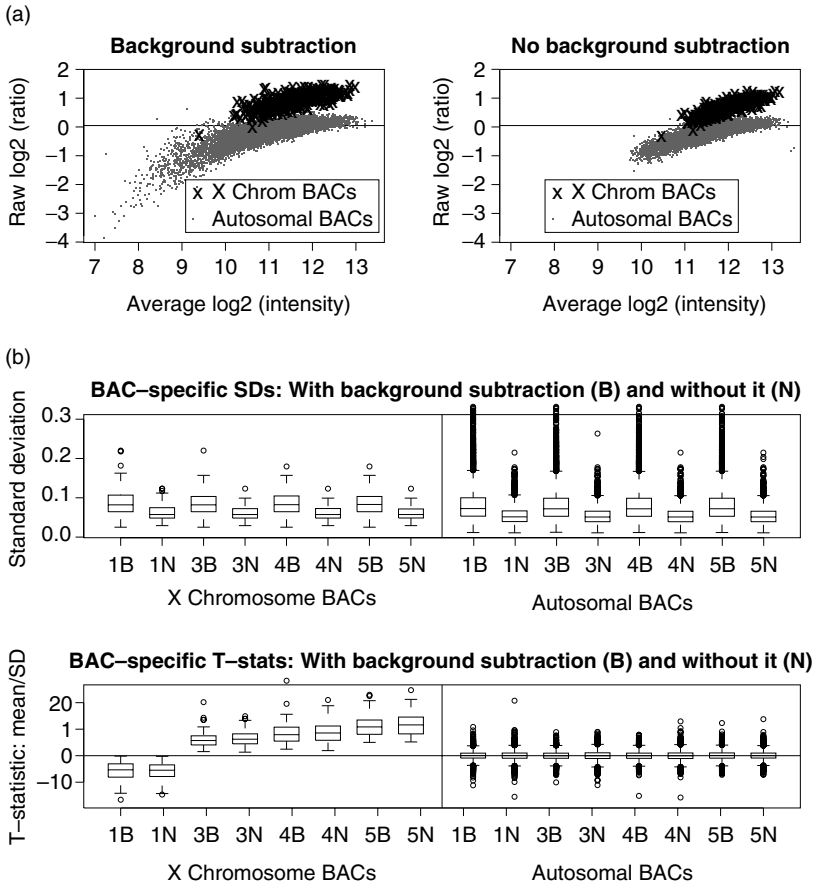


FIGURE 22.1 Comparisons with and without background subtraction. (a) Ratio versus intensity plot of raw unnormalized data for array with 5 copies of X chromosome versus normal female reference DNA with and without background subtraction; (b) boxplots of BAC-specific standard deviations and *t*-test statistics for clones on the autosomes and X chromosomes with and without background subtraction of X-titration experiments from 1X to 5X. The number indicates how many copies of the X chromosome (1, 3, 4, or 5), and the letter (B/N) indicates whether it is background subtraction (B) or no background subtraction (N).

autosomes and X chromosomes, respectively; boxplots of these are shown in Figure 22.1(b) and (c). One can see that standard deviations of X-chromosomal BACs without background subtraction are much smaller than with background subtraction, so that *t*-statistics without background subtraction better discriminate X-chromosomal BACs.

Yang et al.²¹ proposed an alternative background intensity calibration method using morphological opening.²³ Morphological opening is a nonlinear filter that has the effect in a microarray image of removing all spots and replacing them by nearby background values. They showed that this method reduced the variance of

low-intensity spots compared to local background subtraction and achieved better precision compared to no background subtraction. One should consider employing this background correction in software packages such as Spot (<http://experimental.act.cmis.csiro.au/spot>) where it is offered as an option. But for software such as GenePix where the option is not available, it seems from our experience and others¹³ that ratios without background subtraction can have substantially less variability than those with subtraction of local background estimates.

22.2.2 LOCATION ADJUSTMENT

A typical first step in the normalization of array-based measures is a location adjustment. Two basic strategies for location adjustment are in popular use: global location adjustment¹¹ and conserved gene location adjustment.¹² They can be applied either at the array level or the print-tip within array level. Global location adjustment relies on an assumption that most genes are not differentially expressed between the two channels or that as many genes are upregulated as are downregulated. In array-CGH, this assumption means that most clones should not have differential copy numbers between experimental and reference samples or gains and losses of clones on the array should be somewhat balanced. To circumvent this assumption, control gene location adjustment identifies a subset of “housekeeping” genes that are invariable in intensities between the two channels and uses the subset to normalize the whole array. Currently used global and control gene location adjustment methods will be discussed in turn.

The simplest global location adjustment is performed by subtracting the median of all ratios on the array from each individual value. But in many cases, a dye bias appears dependent on the spot intensity so that ratios may exhibit a nonlinear relationship with overall intensity levels. For example, the left panel of Figure 22.2 shows typical *M/A* plots of ratios versus intensities for data before location adjustment for two different print-tips from two arrays. Print-tip A shows a clear relationship between ratios and intensity, in particular at low intensity levels. In contrast, this type of relationship is not observed for print-tip B. To account for the possible systematic differences among print-tips, one can fit a curve for each print-tip separately to capture the trend of ratios as a function of intensities as shown by the solid lines in Figure 22.2. One then normalizes the ratios by subtracting the fitted values from the observed ratios. The curves are obtained using *loess*, a local polynomial regression smoother²⁴ available in the statistical packages R²⁵ and S-PLUS.²⁶ The user must specify the fraction of data used for the locally linear fitting, where larger values produce smoother fitted curves. We determined the value 0.65 empirically using data from a series of normal versus normal arrays. We found that this value was reasonably sensitive for capturing the overall curvature in the data and yet fairly robust to outliers at the high and low intensity levels. An added benefit of print-tip specific location adjustment is the correction for uneven hybridization across the chip. The right panel of Figure 22.2 shows the effect of the *loess* location adjustment applied to data within print-tips. The dependence on intensity, in particular at low or high levels, is removed, and the solid lines fall near the horizontals at 0.

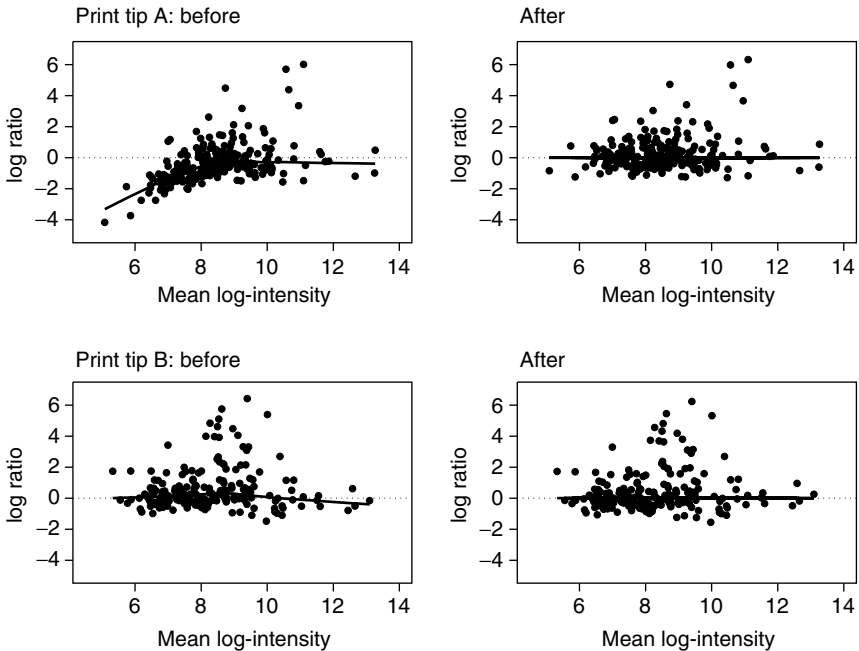


FIGURE 22.2 Within print-tip *loess* curves for data before and after transformation for print-tip #13(A) on array 14 and print-tip #2(B) on array 8, respectively, from the mouse experiment.

For control gene location adjustment, Tseng et al.¹² recommended a rank invariance selection procedure to select a subset of clones that exhibit little variability between the two channels. Their procedure works as follows. The reference and experimental intensities of each clone on the array are separately ranked from smallest to largest based on the intensity level. If, for a given clone, the ranks of red and green intensities differ by more than a threshold value d , which could be the fraction of the data, or if the rank of the average intensity is among the highest or lowest l ranks, then the clone is “discarded”. This process is repeated, beginning with the clones remaining from the previous step, until the set of remaining clones no longer changes. The final set of clones chosen by this procedure is deemed invariant and used to normalize the full set of clones. We applied this procedure to the set of X-titration array-CGH data and noted a couple of difficulties. First, the choice of the thresholds d and l had a substantial impact on the final invariant set. There is no guideline for choosing d and l . Second, the range of the intensities of clones in the invariant set was rather limited, and extrapolation beyond the range was problematic because of the nonlinear dependence of the ratios on intensities, especially at the high and low ends. These problems would be exacerbated on smaller amounts of data, for example, when performing print-tip specific normalization.

To account for the uncertainty in control gene location adjustment, Reilly et al.¹³ proposed a Bayesian framework, treating whether a gene is a control gene as a

parameter and estimating it through Gibbs sampling techniques. While this is appealing, the distributional assumptions imposed on the ratios and parameters in this method are difficult to verify. Huang et al.²⁷ recently proposed semiparametric modeling to relax the distributional assumptions required for this type of location adjustment.

One could consider a combined model for normalization and analysis. For example, the *loess* smoother that was used in the location adjustment earlier could be considered a nonparametric component in the model, whereas the effects of covariates, including treatment and dye, on the ratios could be modeled using linear regression models. The advantages of joint modeling lie in its natural incorporation of uncertainty due to location adjustment and explication of the assumptions made in the normalization step. However, such modeling typically is computationally intensive and may not be able to accommodate downstream multiple comparison procedures or clustering algorithms. Given the current evolving state of microarray technology, we remain skeptical of methods that incorporate normalization or location adjustment as part of the analysis model. These methods may become outdated as the technology improves. We recommend keeping normalization separate from analysis. We think it is also important to provide biologists with the normalized data in addition to the analysis results. The uncertainty in the location adjustment and the measurement of the spot is similar to the classical measurement error problem in statistics, for which many methods have been developed.²⁸ One may incorporate a quality measure, for example variance of the spot intensity or uncertainty of the location adjustment for the gene in the analysis step. Further development toward this direction may be useful in accounting for the location adjustment process in the data analysis.

We end this discussion with a note that we had similar experiences analyzing the cDNA mouse array data. For the same mouse data we also assessed the effect of three global location adjustment methods (median, *loess* using all ratios on the array, and print-tip specific *loess*) in the subsequent analysis for identifying genes that are differentially expressed among the six genetically identical mice; the analysis plan is outlined in Section 22.3. [Figure 22.3](#) shows that correlations of Bonferroni-adjusted $-\log_{10} p$ -values for genes from the three different methods were all greater than 0.9. It shows that even print-tip *loess* did not markedly alter the data for this experiment. This is not always the case, however. Dudoit et al.⁸ discussed an experiment where several print-tips had very different *loess* curves from the rest. In light of Dudoit's findings, it seems prudent to perform a print-tip specific *loess* location adjustment.

22.2.3 WITHIN- AND INTERARRAY SCALE ADJUSTMENT

It is often observed that scales of gene measurements vary among print-tips or arrays, which may be due to technological processes such as uneven hybridization or different batches or laboratories. This type of variation should be normalized to ensure comparable scales across print-tips or arrays. However, expected variation due to different genes or samples should not be normalized. In this section we outline a robust scale adjustment procedure and provide guidelines and exploratory data analyses for deciding when to use it.

The top left panel of [Figure 22.4](#) shows the boxplots of ratios for 16 print-tips of an array from the cDNA mouse data after print-tip specific location adjustment. The

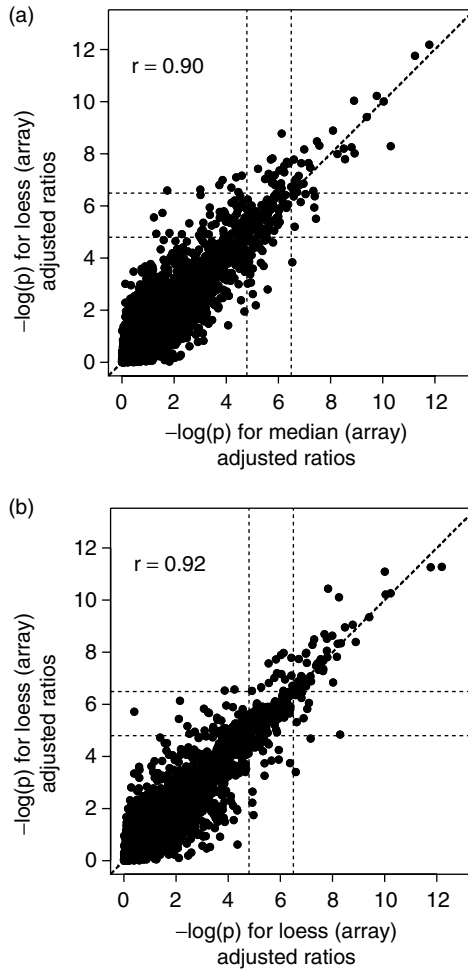


FIGURE 22.3 Scatterplots of Bonferroni-adjusted of $-\log$ base 10 p -values for the pairwise comparisons of normalization methods: (a) global median versus global *loess*, (b) global *loess* versus print-tip-specific *loess*. Dashed lines indicate p -values $0.001/3088$, and $0.05/3088$.

spread of the location-adjusted log ratios varies among the print-tips, although not in any systematic fashion. Given that there may be differentially expressed genes and outliers, a robust measure for the scale is preferable,¹¹ given by

$$a_i = \frac{\text{MAD}_i}{(\prod_{i=1}^I \text{MAD}_i)^{1/I}}$$

where MAD_i is the median absolute deviation for the i th print-tip for $i = 1, \dots, I$. Scale adjustment is performed by dividing each location-adjusted ratio by a_i specific to the print-tip for that ratio.

In addition to varying scales across print-tips within an array, varying scales across arrays in this experiment were also observed, and the variation was of a

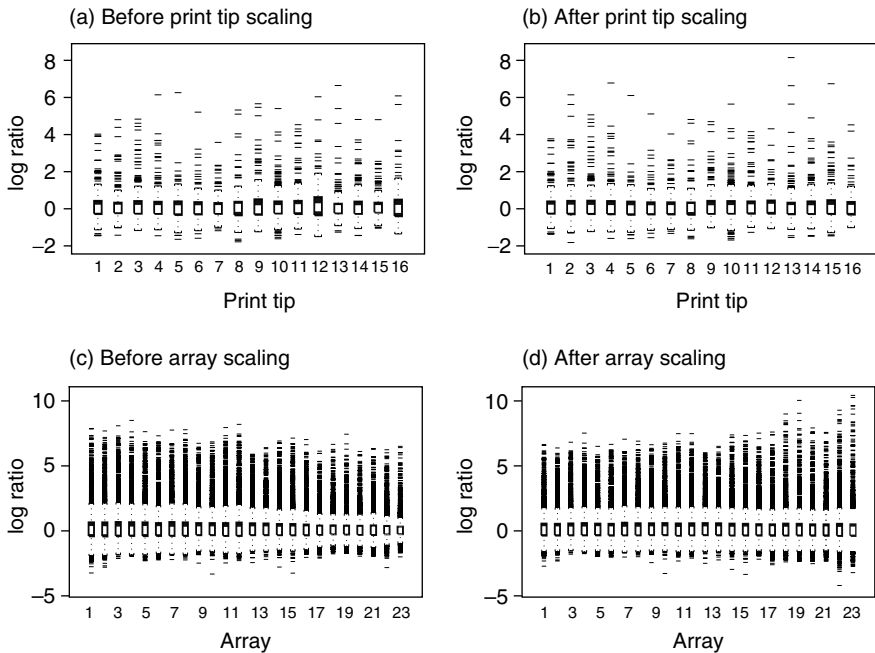


FIGURE 22.4 Boxplots for log ratios from 16 print-tips on array 19 from the mouse experiment before (a) and after (b) scale adjustment. Figure 22.4(c) and Figure 22.4(d) are boxplots for log ratios for the 24 arrays from the mouse experiment before (left) and after (right) MAD array adjustment. All log ratios have been normalized by *loess* within print-tip.

much more systematic nature. The bottom left panel of Figure 22.4 shows the systematic decrease in variability across the 24 arrays after print-tip specific *loess*, which corresponded to the sequential order of mice from 1 to 6 (4 replicates per mouse). The lengths of the interquartile ranges varied from 0.52 to 1.03 with an increasing trend from array 1 to 24. There is no known technical or biological explanation for such fluctuation beyond that the arrays were hybridized sequentially from 1 to 24. We performed the scale adjustment so that the MADs of all arrays are the same and set to be the median of the 24 MADs corresponding to the 24 arrays. After this adjustment, the lengths of the interquartile ranges varied from 0.82 to 0.92.

It was of interest to see how the scale adjustment affected the identification of genes with significant variability in expression levels among the 6 mice. For all comparisons the location-adjustment based on *loess* within print-tips was performed. The correlation of $-\log_{10}p$ -values between methods with and without print-tip scale adjustment appeared to be highly correlated with correlation coefficient 0.92. However the correlation coefficient between $-\log_{10}p$ -values from analyses with an additional scale array adjustment and no additional array adjustment dropped to 0.68. Table 22.1 shows a 3×3 table of the number of genes falling into each of the categories: $p\text{-value} \leq 0.001$, $0.001 < p\text{-value} \leq 0.05$, and $p\text{-value} > 0.05$, where the p -values have been multiplied by the number of genes, 3088, as a Bonferroni

TABLE 22.1
Number of Genes Achieving Each Level of Significance Following *loess* within Print-Tip Location Adjustment with and without Scaling of Arrays^a

Not Scaled	Scaled			
	$p \leq 0.001$	$0.001 < p \leq 0.05$	$p > 0.05$	
$p \leq 0.001$	17	27	24	68
$0.001 < p \leq 0.05$	1	10	145	156
$p > 0.05$	8	33	2823	2864
	26	70	2992	3088

^aThe Log Ratio Method was used, and *p*-values have been adjusted using Bonferroni correction for 3088 comparisons.

adjustment. Not scaling the data yields an exorbitantly high number, 224, of genes declared significant at the 0.05 level and 68 at the 0.001 level. Scaling reduces these numbers to 96 and 26, respectively. In other words, there are an additional 145 genes declared significant at the 0.05 level without the MAD array adjustment that were not significant when the MAD array adjustment was made.

Our next example, however, shows that print-tip or array scale adjustment should not necessarily always be applied. We examined individual scales of the 27 breast tumor array-CGH data and also found considerable differences across arrays (see Figure 22.5). Interestingly, ER– tumors appeared to have greater variation in ratios than ER+ tumors, which is consistent with knowledge that ER– tumors have poorer prognosis and more chromosomal aberrations than ER+ tumors. In this case, the

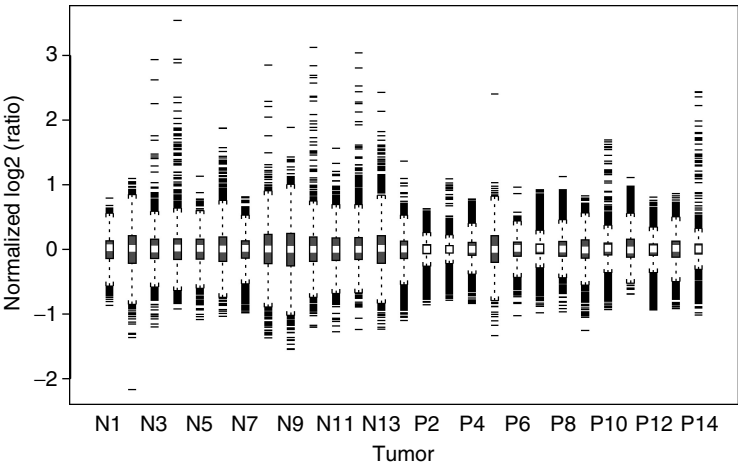


FIGURE 22.5 Boxplots of $\log_2$ ratios for 27 breast tumors using array-CGH data. The first thirteen tumors are ER-negative (labeled N1, . . . , N13), and the last fourteen are ER-positive (P1, . . . , P14). All log ratios have been normalized by *loess* within print-tip.

variable spread among arrays is likely to be real, reflecting the clinical heterogeneity of tumors. Forcing the arrays then to have similar scales may misrepresent the data. This observation emphasizes the importance of a biologically-driven exploratory data analysis for understanding the nature of the scale variation instead of applying a specific correction method blindfolded.

22.3 ANALYSIS

For experiments with single replication, one array per sample, the most common mode of analysis remains multiple two-sample t -tests, along with some multiple comparison correction. The latter is discussed in Section 22.4. There are many variations sought to improve t -statistics to more efficiently handle genes with low overall intensities. A list of recent proposals includes the statistical analysis of microarrays (SAM),²⁹ empirical Bayes,³⁰ Bayesian analyses of microarrays (BAM),³¹ and fully Bayes³² procedures. For data that involve replicates, covariates, or censored outcomes, traditional analysis of variance (ANOVA), generalized linear regression, or the proportional hazards model could be used to assess the effects of variables of interest. The area is rapidly evolving, so many more proposals are expected to appear in the coming years.

An issue that is specific to spotted arrays is whether or not analyses should be based on paired ratios or on absolute intensities. Here we are concerned only with the reference-based design in which a reference is coupled with a sample of interest on each spot. The intensity levels can be obtained by the back transformation (22.1). Intensity-based methods are more easily adapted to study designs other than the reference design, although the reference design remains the most practical design. The efficiency loss of the reference design compared to others such as circular ones is minor when biological variation is large.³³ We will use the cDNA microarray mouse dataset to compare the ratio and absolute intensity level-based methods for identifying significant genes. Analysis of variance (ANOVA) is used for both approaches.

22.3.1 ANOVA FOR RATIOS

For the ratio-based method, let m_{gijk} denote the k th replication of the ratio of the normalized expression level to the reference matched expression level for the g th gene of the i th mouse under dye-color j for $g = 1, \dots, G$, $i = 1, \dots, 6$, $j = 1, 2$, and $k = 1, 2$. For this experiment, $G = 3088$ genes. Then the underlying model for the ANOVA is given by:

$$m_{gijk} = \mu_g + D_{gj} + V_{gi} + \varepsilon_{gijk} \quad (22.2)$$

where μ_g denotes the grand mean of g^{th} gene, D_{gj} denotes the dye effect, V_{gi} denotes the mouse effect, and the ε_{gijk} 's follow independent and identical (i.i.d.) Gaussian distributions with zero mean and constant variance for $g = 1, \dots, G$. Note that there is no effect of array for the 24 arrays because there is no replication of observations within arrays. For each gene model, all effects in (22.2) are fixed with the degrees of freedom given in [Table 22.2](#).

As outlined in Pritchard et al.,²⁰ the model is fit by performing individual regressions for the G genes. Because this is a balanced design, the solution to the parameters

TABLE 22.2
Degrees of Freedom for Each of the 3088
Gene Models Using the Ratio-Based Method

Effect	df
Intercept	1
Dye	1
Mouse	5
Error	17
Total	24

can be obtained explicitly. For each gene an F-statistic, F_g , is computed as the ratio of the mean square for the mouse effect to the residual mean square and compared to the $\mathcal{F}$ distribution with 5 and 17 degrees of freedom to assess significance.

22.3.2 ANOVA FOR INTENSITIES

For absolute expression levels, we adopt the two-stage approach of Wolfinger et al.¹⁵ At the first stage a main effects model is fit to the complete set of G genes. Wolfinger et al.¹⁵ refer to this stage as the normalization model because it is the normalization they use for their experiment, but to avoid confusion with our own normalization procedures outlined in Section 22.2, we will refer to it as the Stage I model. At the second stage, separate gene models are fit to the residuals from the fits obtained from the Stage I model. With no missing data, as is the case here, the two-stage approach is equivalent to the full main effects and interaction model proposed by Kerr et al.¹⁰ Due to the large number of effects, we fit the model in two stages to reduce the computation burden so that standard statistical packages may be used.

Let y_{gijkl} denote the normalized absolute expression level for the k th replication of the g th gene from the i th variety for the j th dye on the l th array for $g = 1, \dots, G$, $i = 1, \dots, 7$, $j = 1, 2$, and $k = 1, 2$, $l = 1, 2, \dots, 24$. Here, variety refers to the collection of six mice and the reference sample. The six mice have two replications for each dye, with each observation on a separate array, whereas the reference sample has twelve replications for each dye, with one observation on each array. Therefore, the design is unbalanced. The Stage I model is given by

$$y_{gijkl} = \mu + D_j + V_i + A_l + \varepsilon_{gijkl}^1 \quad (22.3)$$

where μ denotes the grand mean, D_j denotes the dye effect, V_i denotes the variety effect, A_l denotes the array effect, and the ε_{gijkl}^1 are error terms assumed i.i.d. normally distributed random variables with zero mean and constant variance. Let d_{gijkl} denote the residuals from the fit of (22.3) to the set of 3088 genes from the 6 mice and one reference sample. Then the Stage II model assumes for each gene g that

$$d_{gijkl} = \mu_g + D_{gj} + V_{gi} + A_{gl} + \varepsilon_{gijkl}^2 \quad (22.4)$$

where μ_g is the grand mean, D_{gj} is the dye effect, V_{gl} is the variety effect, A_{gl} is the array effect, and the ε_{gijkl}^2 are i.i.d. zero mean constant variance Gaussian distributions. As the Stage II model is fit separately for each gene, the dye, variety, and array effects are essentially the interactions between dye and gene, variety and gene, and spot effect, in the full model as described in Kerr et al.¹⁰ The residual variance also allows for dependence on the gene. The Stage I and II models can both be fit using SAS Proc Mixed.³⁴ All effects in (22.3) and (22.4) are fixed with degrees of freedom given in Table 22.3.

The F-statistic F_g is computed from the Stage II model as the mean square for variety (excluding the reference variety) divided by the residual mean square and is compared against the $\mathcal{F}$ distribution with 5 and 17 degrees of freedom to assess significance. In practice the F-statistic is computed by defining a matrix L of linearly independent contrasts that reflect all pairwise comparisons between any two varieties excluding the reference sample (see Littell et al.,³⁵ page 482).

22.3.3 COMPARISON OF ANOVA METHODS

The ratio- and intensity-based ANOVA methods may be compared in terms of the genes declared significant. Figure 22.6 shows the Bonferroni-corrected $-\log_{10} p$ -values from both methods. The Pearson correlation coefficient is extremely high, and the p -values are highly collinear, implying that the two methods give virtually identical results. Our initial explanation for the striking similarity was that the variability of the reference sample was much less than that of the individual mice, so that division by expression levels of the reference sample was equivalent to division by a constant. However, examination of the variability in the reference sample revealed that it was of similar magnitude to that of individual mice. Leaving out the reference sample results in little overlap between the intensity level-based and ratio-based approaches; the correlation drops to 0.13. It therefore appears that the similarity was really due to the spot or array effects in Equation (22.4) being adjusted for by the intensity model in the same way as by taking the ratio. We refitted the intensity-based model ignoring the spot effect and found virtually no overlap between the two approaches (correlation coefficient 0.14). Therefore, we concluded that inclusion of the spot effect is essential in fitting the intensity-based model and that under this assumption the two different approaches can be expected to give similar results.

TABLE 22.3
Degrees of Freedom for the Stage I and II Models Using
the Expression-Based Approach

Effect	Df (Stage I)	Df (Stage II)
Intercept	1	1
Dye	1	1
Variety	6	6
Array	23	23
Error	148193	17
Total	148224	48

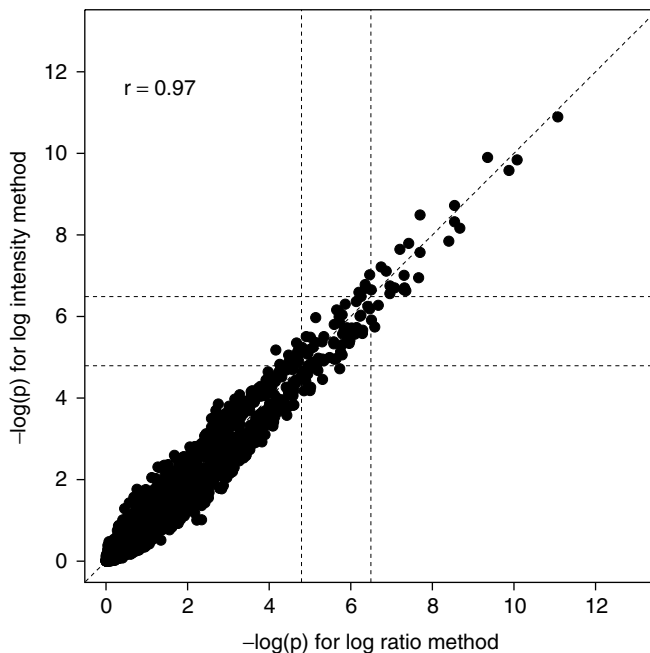


FIGURE 22.6 Comparison of p -values from the ratio- versus intensity-based method; data were normalized using the within-print-tip and MAD withinarray normalization scheme. Dashed lines indicate p -values 0.001/3088 and 0.05/3088.

22.4 MULTIPLE COMPARISONS

Microarray experiments generate large multiplicity problems in which thousands of hypotheses, typically one per gene, are tested simultaneously. For a test, two types of errors are possible. The first is known as the type I error rate, or false positive rate, and is the fraction of clones declared significantly different from the null hypothesis that are not truly different. The second is the type II error rate, or false negative rate, and is the fraction of clones not declared different from the null hypothesis that actually are different. In a typical testing situation, one controls the type I error rate to fall below a prespecified value, say 0.05, while minimizing the type II error rate. When many hypotheses are tested, each with a specified type I error probability, the chance of committing some type I errors increases, often sharply with the number of hypotheses tested.³⁶ Ge et al.¹⁹ recently provided a comprehensive review of various multiple comparison procedures, including the family-wise error rate (FWER) procedures of Westfall and Young,¹⁶ the false discovery rate of Benjamini and Hochberg,¹⁷ and the positive false discovery rate of Storey.³⁷ In general, the FWER procedures were found to be most stringent, the FDR most relaxed, and the pFDR somewhere in between. They also compared these multiple comparison procedures using two cDNA microarray datasets. They observed that when just a few genes are expected to be differentially expressed, the FWER-based procedures appeared to have better power to detect these genes, while when many genes are expected to be

differentially expressed, the FDR- or pFDR-based procedures had better power. In this section we will assess how various multiple comparison procedures perform for our array-CGH data.

A main difference between array-CGH data and cDNA array data lies in the dependence structure. It is commonly believed that genes or clones tend to act concordantly in the cells. Therefore the probability of a clone showing an aberration increases, sometimes substantially, if the neighboring clones show aberrations. Because array-CGH data measure clones along the chromosomes, we expect there will be dependence induced between clones by physical proximity on a chromosome. How this dependence affects the performance of multiple comparison procedures is largely unknown. We investigate this issue through the application of several multiple comparison procedures to an analysis of array-CGH data from a set of 27 breast tumors.

We consider five multiple comparison procedures: 1) Bonferroni; 2) Westfall and Young's¹⁶ step-down minP (the fast implementation of Ge et al.);¹⁹ 3) Benjamini and Yekutieli's³⁸ estimator of the false discovery rate; 4) Storey and Tibshirani's¹⁸ resampling-based false discovery rate; and 5) Storey and Tibshirani's¹⁸ resampling-based positive false discovery rate. Other variations of multiple comparison procedures are available, but some rely on strong assumptions to control the type I error rates. For example, a widely used false discovery rate procedure proposed by Benjamini and Hochberg¹⁷ controls FDR only under positive regression dependency. Such assumptions in general are difficult to verify in high-dimensional data. Here we will only compare procedures that are derived under more general dependence assumptions. All procedures will be based on raw or unadjusted p -values of the clones rather than the test statistics, as the null distributions for all the clones are not necessarily comparable with each other. Because it is difficult to individually verify whether the test statistic follows a known parametric distribution for each clone, we use a permutation procedure with 1,000,000 permuted datasets to estimate the raw p -value for each clone. In what follows we will briefly describe each of the 5 procedures used in the comparison; interested readers may refer to Ge et al.¹⁹ for further description of the methods.

The Bonferroni procedure is perhaps the most well-known procedure in multiple testing. The adjusted p -values are calculated by multiplying the raw p -values by the total number of clones, G , then setting those greater than the upper bound of 1 to 1. It provides strong control of the family-wise error rate at level α . Recall that the FWER is the probability of having at least one type I error among the m hypothesis tests and that strong control is control under every combination of true and false hypotheses.

Westfall and Young's step-down minP procedure also provides strong control of the FWER. Ranking the raw p -values from the smallest to the largest, the term step-down refers to its process of calculating the adjusted p -values from the most significant clone to the least significant one. The procedure operates conceptually by repeating the following steps.

1. Permute the group indices to create a new permuted dataset and compute raw p -values for all clones for that permuted dataset.
2. For the k th order clone calculate its minP value by taking the minimum of the raw p -values from step 1 for the clones that had higher rank in the original data than the k th order.

These steps should be performed for all possible unique permuted datasets, if it is computationally feasible or else for a random sample of at least 100,000, preferably 1,000,000, such sets. The adjusted p -value for a clone is defined to be the fraction of permuted datasets in which the minP value is less than or equal to the raw p -value. Monotonicity constraints are enforced by taking a step-down process in which the adjusted p -value for the k th order clone is set to be the larger of adjusted p -values of its own and the lower ranking ones. The minP procedure may potentially gain power when the dependence among the clones is strong.

Rather than controlling the FWER, procedures have been proposed to control the FDR, the expected proportion of type I or false positive errors among the rejected hypotheses. Different methods for handling the case of 0 rejected hypotheses lead to different definitions. Benjamini and Hochberg¹⁷ set the FDR equal to 0 under 0 rejected hypotheses, whereas Storey³⁷ defines positive FDR (pFDR) as the expectation of the proportion of type I errors among the rejected hypotheses conditional that at least one hypothesis is rejected.

Benjamini and Yekutieli³⁸ proposed a simple step-up procedure for strong control of the FDR under arbitrary dependence. Ordering the raw p -values from smallest to largest, they first multiply the raw p -values by the total number of clones G and $C = \sum_{g=1}^G 1/g$, a correction term to account for the general dependency among the null hypotheses, and then divide by the rank of the clone among all the clones. The adjusted p -value for the k th order p -value of the clone is simply the minimum of the p -values of all clones with higher rank than the k th order. At any step, if a p -value exceeds 1.0, it is reset to 1.0. We will refer to Benjamini and Yekutieli's procedure henceforth as the BY procedure.

A resampling-based procedure for estimating the FDR and pFDR is given by Storey and Tibshirani.¹⁸ The estimator was derived under a general dependence assumption that involves certain ergodic conditions. The assumption regarding the dependence is more restrictive than either minP or the BY procedure. Whether the array data satisfy the dependence conditions remains to be verified. Storey and Tibshirani¹⁸ provides an algorithm that is based on the values of test statistics. We modified it following partially the Ge et al.'s¹⁹ fast minP algorithm, so that the FDR calculations will also be based on the raw p -values like other multiple comparison procedures. The resampling-based pFDR follows essentially the FDR procedure but divides the FDR by an estimate of the probability of one or more hypothesis being rejected. The q -values are estimated by enforcing step-down monotonicity on the pFDR estimates.

Figure 22.7 displays results of the procedures that control for the FWER (top panel) and the FDR (bottom panel) applied to the array-CGH data. The plots show the FWER adjusted p -value, q -value, or FDR estimates on the y-axis versus the rank of the raw p -value on the x-axis. One can see that the minP procedure generally yields smaller adjusted p -values than the Bonferroni procedure. However, when we focus on the small p -values, such as those less than 0.1 where the main interest usually lies, the difference between the two methods is much less. In these data, we find that the minP procedure rejects 6 BACs at level 0.05 and 12 at level 0.10, compared to 5 and 8, respectively, for Bonferroni.

Of the FDR procedures, the BY is clearly the most conservative; see the bottom panel of Figure 22.7. Storey and Tibshirani's FDR estimate is lower than their pFDR

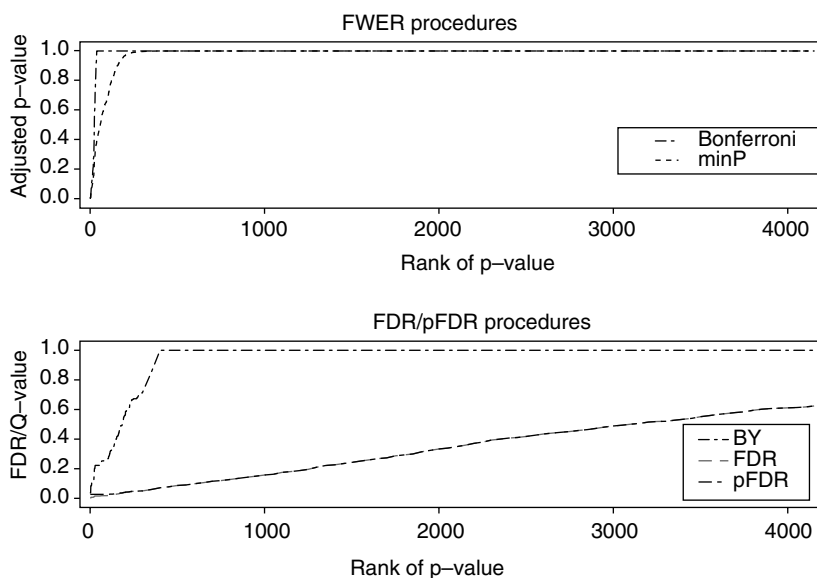


FIGURE 22.7 Comparison of adjusted p -values from various multiple comparisons procedures applied to the breast tumor array-CGH data. (a) Adjusted p -values versus the rank of the clones for procedures controlling FWER (Bonferroni and minP); (b) adjusted p -values versus the rank of the clones for procedures controlling FDR or pFDR (BY, FDR, and pFDR of Storey and Tibshirani).

estimate for about the 200 lowest ranked p -values, after which the two methods essentially coincide. This is expected as the pFDR is equal to the FDR divided by the probability of at least one hypothesis being rejected. This probability term will increase to 1 with the p -value. Using a cutoff of 0.10 for the FDR and pFDR, we find that BY rejects only the top 9 ranked BACs with an expectation of just under 1 false positive, while Storey and Tibshirani's FDR and pFDR estimators both reject 607 with an expectation of 60–61 false positives.

One may also consider multiple comparison procedures based on observed test statistics rather than raw p -values. The computational burden is higher for the latter, because it involves calculating the raw p -values not only for the original data but also for all permuted datasets. In addition, the sample size of a dataset may be too small to get sufficient precision in the raw p -values. In these situations, procedures that are based on the observed test statistics may be preferred. However, care must be taken to make sure that the null distributions of the test statistics are reasonably comparable with each other for all clones. A simple way to verify this assumption is to examine whether the ranks of the raw p -values and absolute test statistics are reasonably identical. For the breast tumor data, the ranks of raw p -values and absolute test statistics are roughly in a linear trend, but inconsistencies are observed. In this case, we would recommend the procedures based on the raw p -values rather than the test statistics.

It is apparent that for our breast tumor data the FWER procedures and the BY procedure for the FDR are drastically different from the FDR/pFDR procedure of

Storey and Tibshirani.¹⁸ Our feeling is that the truth is likely somewhere inbetween. While the BY and Storey and Tibshirani procedures control an expectation of the FDR, they do not provide any information on the variability of these estimates. Rather than controlling for the mean of FDR, Lehmann and Romano³⁹ proposed to control the probability of the false discovery proportion exceeding a prespecified level. They also proposed to generalize the control of the standard FWER to control the probability of making k or more errors, a concept that is quite useful when biologists are looking for more clues and are willing to tolerate a few more errors. Hopefully the increased interest in multiple testing brought about by the explosion of biological data will lead to more advances in this area.

22.5 FINAL REMARKS

We have summarized and provided practical guidance for the analyses of spotted microarrays based on our own experience. We expect the technology to improve, hopefully to the point where many of the normalization issues become moot. Until then exploratory data analysis remains key to choosing the appropriate normalization procedure, and one should not expect one type of procedure to be appropriate for all arrays. Statisticians are still on a learning curve with respect to statistical methods for analyzing high-dimensional data. We expect many more advances in this exciting and challenging area. The programs used to generate the results in this chapter are available from the authors upon request. Many of these methods and procedures are also available through the R Bioconductor project at Web site <http://www.bioconductor.org/>.

ACKNOWLEDGMENT

The authors are grateful to Drs. Peggy Porter and Peter Nelson of the Fred Hutchinson Cancer Research Center for providing the cDNA data on the mouse experiment and the array-CGH data on the X-titration and the breast tumor, respectively. This research is supported in part by NIH grants.

REFERENCES

1. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. *Nat.Genet.* 1999;21:33–37.
2. van't Veer LJ, van de Vijer MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. *Nat.Genet.* 2002;415:530–536.
3. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM. Delineation of prognostic biomarkers in prostate cancer. *Nature* 2001;412:822–826.
4. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* 1998;20:207–211.

5. Pollack JR, Peron CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CG, Jeffrey SS, Botstein D, and Brown PO. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* 1999;23:41–46.
6. Mantripragada KK, Buckley PG, de Stahl TD, Dumanski JP. Genomic microarrays in the spotlight. *Trends Genet.* 2004;20:87–94.
7. Loo LWM, Grove DI, Neal CL, Cousens LA, Schubert EL, Williams EM, Holcomb IL, Delrow JJ, Trask BJ, Hsu L, and Porter P. Array CGH analysis of genomic alterations in breast cancer sub-types. Submitted.
8. Dudoit S, Yang YH, Speed TP, Callow MJ. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 2002;12(1):111–139.
9. Kerr MK, Churchill GA. Experimental design for gene expression microarrays. *Biostat.* 2001;2:183–201.
10. Kerr MK, Afshari CA, Bennett L, Bushel P, Martinez J, Walker NJ, Churchill GA. Statistical analysis of a gene expression microarray experiment with replication. Submitted 2002.
11. Yang YH, Dudoit S, Luu P, Speed TP. *Normalization for cDNA Microarray Data*. SPIE BIOS 2001, San Jose, California.
12. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations, and assessment of gene effects. *Nucleic Acids Res.* 2001;12:2549–2557.
13. Reilly C, Wang C, Rutherford M. A method for normalizing microarrays using genes that are not differentially expressed. *J. Am. Stat. Assoc.* 2003;98:868–878.
14. Breslow NE, Day NE. *Statistical Methods in Cancer Research*, vol. 1, *The Analysis of Case-Control Studies*. World Health Organization, 1980.
15. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comp. Biol.* 2001;8:625–637.
16. Westfall PH, Young SS. Resampling-based multiple testing: *Examples and Methods for p-Value Adjustment*. New York: Wiley, 1993.
17. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J.R. Statist. Soc. B.* 1995;57:289–300.
18. Storey JD, Tibshirani R. Estimating false discovery rates under dependence, with applications to DNA microarrays. Technical report, Department of Statistics, Stanford University.
19. Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. *Sociedad de Estadística e Investigación Operativa Test* 2003;12:1–77.
20. Pritchard CC, Hsu L, Delrow J, Nelson PS. Project normal: defining normal variance in mouse gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 2001;98;23:13,266–13,271.
21. Yang YH, Buckley MJ, Speed TP. Analysis of cDNA microarray images. *Briefings Bioinformatics* 2001;2:341–349.
22. GenePix Pro 4.1. Axon Instruments Inc., Copyright 1995–2002.
23. Soille P. *Morphological Image Analysis: Principles and Applications*. Berlin Heidelberg: Springer-Verlag, 1999.
24. Cleveland WS, Grosse E, Shyu WM. Local regression models. In *Statistical Models in S*, eds. J.M. Chambers and T.J. Hastie. Wadsworth & Brooks/Cole Pacific Grove, CA:1992.
25. Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2004. <http://www.R-project.org/>.

26. Insightful Corporation. *S-PLUS 2002 Professional Release 1*. <http://www.insightful.com/splus>.
27. Huang J, Wang D, Zhang CH. A two-way semi-linear model for normalization and significant analysis of cDNA microarray data. The University of Iowa, Department of Statistics and Actuarial Science, Technical Report No. 326, 2004.
28. Little RJA, Rubin DB. *Statistical Analysis With Missing Data*. New York, Chichester: John Wiley & Sons, 2002.
29. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Nat. Acad. Sci. U.S.A.* 2001;98:5116–5121.
30. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* 2001;96:1151–1160.
31. Ishwaran H, Rao JS. Detecting differentially expressed genes in microarrays using Bayesian model selection. *J. Am. Stat. Assoc.* 2003;98:438–455.
32. Ibrahim JG, Chen M-H, Gray RJ. Bayesian models for gene expression with DNA microarray data. *J. Am. Stat. Assoc.* 2002;97:88–99.
33. Kerr MK. Design considerations for efficient and effective microarray studies. *Biometrics* 2003;59:822–828.
34. SAS Institute Inc. *SAS/STAT User's guide, Version 8*. Cary, NC: SAS Institute Inc., 1999.
35. Littell RC, Milliken GA, Stroup WW, Wolfinger RD. *SAS System for Mixed Models*. Cary, NC: SAS Institute Inc., 2000.
36. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. *Stat. Sci.* 2003;71–103.
37. Storey J. The positive false discovery rate: a Bayesian interpretation and the q -value. *Ann. Stat.* 2003;31:2013–2035.
38. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 2001;29:1165–1188.
39. Lehman EL, Romano JP. Generalizations of the familywise error rate. Technical report No 2003-37, Department of Statistics, Stanford University.

23 Statistical Applications Using DNA Microarrays for Cancer Diagnosis and Prognosis

Shigeyuki Matsui

CONTENTS

23.1	Introduction	419
23.2	Gene Expression Profiling for Bladder Cancer	420
23.3	Class Discovery	421
23.3.1	Clustering	421
23.3.2	Alternative Dimension Reduction Techniques	423
23.3.3	Illustration	423
23.4	Gene Identification.....	424
23.4.1	Multiple Testing	424
23.4.2	Model-Based Approach	425
23.4.3	Illustration	425
23.5	Prediction	427
23.5.1	Selection of Genes	427
23.5.2	Development of Prediction Models	428
23.5.3	Estimation of Predictive Accuracy.....	429
23.5.4	Illustration	429
23.6	Software Note	431
	Acknowledgments	432
	References	432

23.1 INTRODUCTION

Cancer is a highly complex disease with somatic mutations and epigenetic changes.¹ Many of these molecular abnormalities, which are specific to individual neoplasms, influence the expression of genes that control the growth of a tumor, invasiveness, metastatic potential, and responsiveness or resistance to treatment. Accordingly, a major focus in clinical oncology is to develop markers for accurate diagnosis and

precise prognostic prediction based on tumor gene expression profiling.² Such markers can be useful in the selection of appropriate therapeutic approaches for individual patients and in the design of future therapeutic trials. The final deciphering of the complete human genome, together with the improvement of high-throughput assays, has initiated the development of such markers based on a large proportion of the genes on a genome. DNA microarrays are a new and promising high-throughput assay that allow the simultaneous measurement of the level of expression for thousands of genes, or even an entire genome, from human tumor samples. For an introductory overview of biological and technical aspects of microarray assays, see Nguyen et al.³

During the past few years many studies directed toward clinical application of DNA microarrays have emerged in cancer research. From these studies we can identify three types of objectives: class discovery, gene identification, and prediction. Class discovery refers to discovering novel subclasses of cancer based on tumor gene expression profiling for a clinically and morphologically similar cancer.^{4–7} For example, Alizadeh et al.⁴ identified two previously unrecognized subgroups of diffuse large B-cell lymphoma by hierarchical clustering of patients based on tumor gene expression profiling, and interestingly, these subgroups had different outcomes after multiagent chemotherapy. The discovered expression-based subclasses have a potential to reflect distinct pathogeneses that cannot be discerned by standard morphological and pathological criteria. Gene identification refers to selecting genes that are differentially expressed across prespecified classes.^{6–11} Hedenfalk et al.⁹ identified different groups of genes expressed by breast tumors with BRCA1 mutations and those with BRCA2 mutations. Prediction corresponds to developing expression-based prediction systems that predict diagnostic classes or outcomes. Expression-based predictors for diagnostic classes have the potential to improve the diagnostic accuracy of current approaches.¹² Golub et al.⁶ developed predictors for classification of acute leukemia arising from lymphoid and myeloid precursors. Examples of expression-based prediction for outcomes include prediction of response after an IL2-based treatment for melanoma¹³ and prediction of survival time after treatment for diffuse large B-cell lymphoma,^{14,15} early-stage lung adenocarcinoma,¹⁶ and metastatic kidney cancer.¹⁷

In this chapter, we shall focus on statistical issues in the analysis of normalized data from cancer clinical studies with microarrays. After introducing microarray data from a bladder cancer study that is used for illustration in Section 23.2, we outline statistical methods for class discovery, gene identification, and prediction in Sections 23.3, 23.4, and 23.5, respectively. A software note appears in Section 23.6.

23.2 GENE EXPRESSION PROFILING FOR BLADDER CANCER

The data in this chapter come from a pretreatment cDNA microarray for bladder cancer. Tumor-biopsy tissues were collected at diagnosis from 55 bladder cancer patients who presented themselves at the Department of Urology, Kyoto University Hospital in Japan after 2000. Patients were selected on the basis of availability of tumor-biopsy tissues without regard to clinical information. Microarray experiments were performed for frozen tissues using a cDNA array that contains printings of

approximately 30,000 oligonucleotides (DNA fragments that correspond to all the known and the majority of predicted genes in the human genome) fabricated by Pacific Edge Biotechnology Limited in New Zealand. Distinct oligonucleotides correspond to distinct genes. In each hybridization, fluorescent cDNA targets were prepared from a tumor mRNA sample and a reference mRNA sample contracted from a pool of cell lines of different cancers. After image analysis and spot filtering, the data were normalized by a locally weighted linear regression method.^{18,19} A chi-square test of the hypothesis that the variance in expression across patients is equal to the median of such variance was performed separately for each gene assuming that the variability in expression of a large part of genes with lower variance is not likely to reflect actual biological variability. Using the significance level of 1% for this test, 6,000 genes with the largest variances were selected for subsequent analyses. Clinical data involve patients' background characteristics, diagnosis information including stage and grade, treatments, and outcomes including response in tumor shrinkage, recurrence, and death. The diagnostic information was confirmed by a pathologist for all the patients. We selected 48 transitional cell carcinomas (TCCs), which are the most prevalent bladder tumors, for the analysis, excluding squamous cell carcinomas (SCCs) and adenocarcinomas. The main focus in this chapter is on correlating gene expression data with important diagnostic variables such as stage and grade.

23.3 CLASS DISCOVERY

This section discusses unsupervised clustering and dimension reduction methods and graphical displays for exploring microarray data. Such methods are useful for class discovery.

23.3.1 CLUSTERING

The goal of clustering is to group together objects (genes or samples) so that expression patterns within a group are more alike than patterns across groups. There are two broad types of clustering algorithms, *hierarchical clustering* and *partitioning clustering*.²⁰

Hierarchical clustering is the most widely applied clustering algorithm for class discovery. It works by producing a series of successively nested clusters, organizing a tree structure or dendrogram for a prespecified matrix of pairwise dissimilarity or distance between objects and distance between clusters that is specified by an average, complete, or single linkage method.²¹ A single partition is obtained by cutting the dendrogram at a particular level. There are two types of hierarchical clustering algorithms: agglomerative and divisive clustering.²⁰ Agglomerative clustering operates by iteratively joining the two closest clusters starting from a singleton cluster, whereas divisive clustering operates by iteratively partitioning clusters starting with the complete set of the data. Divisive algorithms are less commonly used in practice, partly due to their serious computational burden. A major limitation of hierarchical clustering algorithms derives from their sequential nature, which may result in a high risk of clustering on noise. For example, agglomerative algorithms, for which each

merge depends on all previous merges, cannot recover from bad merges that occur at earlier stages. Consequently, the clustering results can poorly reflect structure near the top of the tree, where there are relatively few large clusters. This problem is exacerbated in microarray data where there is a large number of noisy variables. One practical approach is to apply hierarchical methods as a preliminary step followed by partitioning clustering methods described below as a final clean-up step.²²

Partitioning clustering algorithms produce a single collection of nonnested disjoint clusters for a prespecified number of clusters and initial partitioning. The *k*-means clustering,²³ *k*-medoids clustering,²⁴ and self-organizing maps (SOM)^{25,26} are such algorithms that have been applied to microarray data. For a given number of clusters *k* and initial cluster centers, *k*-means clustering partitions the objects so that the sum of squared distances of each object to its closest cluster center is minimized. The *k*-medoids clustering uses medoids instead of centroids for the centers of clusters. SOM is a neural network procedure that can be viewed as a constrained version of *k*-means clustering that forces the cluster centers to lie in a discrete two-dimensional space in which the clusters are sorted according to their degree of similarity. An advantage of partitioning clustering is that, through utilizing the prior information on the number of clusters, they reduce the risk of clustering on noise, a weakness of hierarchical clustering, although one does not typically know the number of clusters and the prior information can be incorrect. Partitioning clustering is less computationally demanding than hierarchical clustering, which is particularly advantageous for clustering thousands of genes. An important practical issue is how to choose the initial partitioning, which can largely impact the final result. Some approaches are to perform other procedures such as hierarchical clustering or to run a partitioning procedure repeatedly for different randomly chosen sets of initial cluster centers. Then one may choose the partition that minimizes the within-cluster sum of square for a given number of clusters.²⁷

Two-way clustering simultaneously clusters genes and samples with a goal of identifying groups of genes involved in multiple biological activities in subsets of samples. A simple two-way clustering could be found by reordering the genes and samples after independently clustering them, such as available in the Eisen software²¹ (also see Section 23.6). More complex methods developed recently include coupled clustering,²⁸ block clustering,²⁹ gene shaving,³⁰ and the plaid model.³¹

Objective assessment of validity of clustering is particularly important because clustering algorithms can always produce clusters even on random data. A number of estimation methods for determining the number of clusters have been proposed in statistical literature; see Milligan and Cooper³² and Gordon²⁰ for a comprehensive review. Dudoit and Fridlyand³³ reported that a prediction-based resampling method on the reproducibility of cluster assignment was more accurate and robust than methods based on measures of cohesiveness within clusters including silhouette widths,²⁴ the gap statistic,³⁴ and others.^{35,36} Other prediction-based resampling methods include a stability-based resampling method³⁷ and a jackknife-type method.²² Model-based partitioning clustering with underlying mixture models would also be useful for clustering genes to compare to competing clustering results with different numbers of clusters.^{38–40}

Another aspect of cluster validation is the assessment of stability or reproducibility of individual clusters. In data perturbation methods, artificial random noise is added to the observed data; the data are reclustered; and the difference with the original clustering results is evaluated.^{41,42} McShane et al.⁴¹ proposed two reproducibility indices, robustness (R) index and discrepancy (D) index. The R -index measures the proportion of pairs of objects within a cluster for which the members of the pair remain together in the reclustered perturbed data. The D -index compares the number of discrepancies in an original cluster to a best-matching cluster, defined as the one having the greatest number of objects in common with the original cluster.

23.3.2 ALTERNATIVE DIMENSION REDUCTION TECHNIQUES

Principal component analysis (PCA), sometimes referred to as singular value decomposition, and multidimensional scaling (MDS) are prototype dimension reduction techniques that have been widely used in microarray studies.^{5,8,43,44} A drawback to PCA is that summary variables, which are the orthogonal linear combinations of potentially large numbers of genes showing the greatest variability across samples, do not necessarily have a clear biological interpretation. In addition, there is generally no guarantee that the data will cluster along the dimensions identified by these techniques when there are a large number of noisy variables.

23.3.3 ILLUSTRATION

We performed hierarchical clustering of 48 bladder cancer tumors based on their expression profiles of the 6,000 genes that passed a gene filtering step; see Section 23.2. The analysis was performed using BRB ArrayTools (see Section 23.6). For a dissimilarity metric, one minus the Pearson correlation was used after median centering for each gene. Figure 23.1 shows dendrograms when the average linkage method was used. The same six clusters were also found for the complete linkage method. Table 23.1 summarizes the indices of robustness and discrepancy in terms of omissions and additions proposed by McShane et al.⁴¹ for the six clusters. One

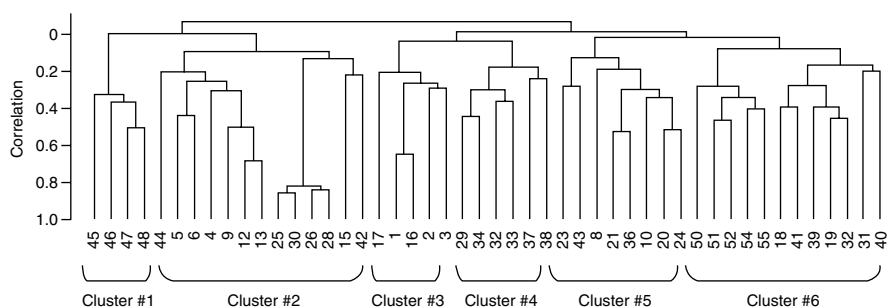


FIGURE 23.1 A dendrogram resulting from hierarchical clustering of gene expression profiles for 48 bladder tumors (coded from 1 through 48). One minus Pearson correlation was used as the dissimilarity metric after median centering of log-ratio of expression for each gene. The average linkage method was used. Six clusters were identified.

TABLE 23.1

Reproducibility Indices for Each of the Six Clusters Found by Hierarchical Clustering with the Average Linkage Method for 100 Perturbed Datasets

Cluster	Number of samples in cluster	Robustness	Omission	Addition
Cluster #1	4	1.00	0.00	0.00
Cluster #2	13	0.97	0.18	0.03
Cluster #3	5	0.99	0.02	0.40
Cluster #4	6	0.99	0.03	0.24
Cluster #5	8	0.91	0.42	0.03
Cluster #6	12	0.98	0.23	0.02

Robustness refers to the averaged proportion of pairs of objects within a cluster for which the members of the pair remain together in the reclustered perturbed data over 100 perturbed datasets. Omission refers to the averaged number of objects in the original cluster that are not in the best match perturbed cluster. Addition refers to the averaged number of objects in the best match cluster that are not in the original cluster.

hundred perturbed data using Gaussian random noise with a median variance of the log-ratios of gene expression were used to calculate these indices. Omissions refer to the number of objects in the original cluster that are not in the best match perturbed cluster, and additions refer to the number of objects in the best match cluster that are not in the original cluster. Entries in the column are the averaged number of omissions and additions over the 100 perturbed datasets. The robustness and discrepancy indices indicated high reproducibility for the six clusters. Similar values of the indices were obtained when the number of clusters was set to be four or eight. With respect to association with clinical information, eight out of nine tumors with the worst types, stage ≥ 3 and grade = 3, were in cluster #2.

23.4 GENE IDENTIFICATION

Screening for differentially expressed genes between prespecified classes is often a primary aim in clinical applications. Typically separate statistical tests are made for each gene such as a two sample *t*-test for comparing gene expression levels between two classes. Examination of many hypotheses sharply increases the number of false positives.

23.4.1 MULTIPLE TESTING

One of the earliest and most commonly used approaches for multiple testing, the Bonferroni procedure controls the *family-wise* error. If m genes are examined with unadjusted p -values p_j for gene j , the *Bonferroni-adjusted* p -value is the minimum of mp_j and 1. The Bonferroni procedure can be too conservative when m is large. Less conservative procedures can be obtained by sequential adjustments depending on the order of the observed p -values.^{45,46} Westfall and Young⁴⁷ proposed a multivariate permutation method that takes correlation between genes into account and hence is less conservative. For a single permutation of the class labels, unadjusted univariate p -values are ordered, $p_{(1)}^* \leq p_{(2)}^* \leq \dots \leq p_{(m)}^*$. For a large number of

permutations, B , an estimate of the *single-step min P adjusted p-value* for gene j , is obtained by the number of permutations where $p^*_{(1)} \leq p_j$ divided by B . Slightly less conservative procedures can be obtained using stepwise modifications.⁴⁷ Note that the null distribution satisfies the complete null hypothesis that the null hypothesis is true for all genes, otherwise termed as weak control. Strong control refers to control under any combination of true and false null hypotheses.⁴⁷

Other approaches to multiple testing control the false discovery rate (FDR), defined as the expected proportion of false positives among the genes declared significant.⁴⁸ The FDR can be estimated by a multivariate permutation procedure. For a given cutoff c on a continuous measure of differential expression p^* , an estimate of FDR is given by the average of h that satisfies $p^*_{(h)} \leq c$ and $p^*_{(h+1)} > c$ over all permutations divided by the number of p^* 's less than c . One chooses a cutoff for p^* yielding an acceptable FDR. Other references for the FDR are given in.^{49–54}

23.4.2 MODEL-BASED APPROACH

Model-based approaches can be suitable for some situations. If the number of samples is small for each class under comparison, then individual variances cannot be accurately estimated, reducing the statistical power of the multiple testing procedure. Hierarchical multilevel modeling is a more efficient approach in which gene-specific summarizes such as mean and variance of expression follow a mixture distribution.^{52,55–61} ANOVA-type modeling for the channel-specific intensities for each gene in each array is another way to approach the problem.^{62–65} Model-based approaches are also useful for a multiclass test in which it is necessary to adjust for classes and/or to evaluate possible interactions among classes. A simple method on a one-gene-at-a-time basis is illustrated below for this type of analysis.

23.4.3 ILLUSTRATION

Identifying genes differentially expressed for important diagnostic classes such as stage and grade is an important objective in the bladder cancer study described in Section 23.2 for a better understanding of the biological differences between the diagnostic classes and for exploring expression-based markers for accurate diagnosis. For a single gene, let Y_j be the log-ratio of gene expression measurement for sample j ($j = 1, \dots, 48$). Let stage_j be a binary variable that takes 1 if histopathological stage of sample $j \geq \text{T1}$ (invasive) and zero otherwise (superficial), and let grade_j be a binary variable that takes 1 if the grade of sample j is 3 (high grade) and zero otherwise, i.e., $\text{grade} < 3$ (low grade). We assumed the linear model,

$$Y_j = \beta_0 + \beta_1 (\text{stage}_j) + \beta_2 (\text{grade}_j) + \beta_3 (\text{stage}_j \times \text{grade}_j) + \varepsilon_j, \quad (23.1)$$

where β_0 , β_1 , β_2 , and β_3 represent parameters. This model accounts for possible interaction between stage and grade. We assumed the error term ε_j to have mean zero and to be independent of each other. The model (23.1) is similar to the ANOVA-type model for two-class comparison in Thomas et al.⁶⁴ after ignoring the preprocessing step. We fitted model (23.1) separately for each gene. Gene selection was

based on statistical significance of an F test for evaluating goodness of fit of the model, which is a test of the null hypothesis that the mean expression levels are identical across the four groups that correspond to four combinations of two binary classes, stage and grade. The SAS procedure `proc reg` was used to fit the model.

Table 23.2 shows an estimate of FDR for each of several P -value cut-offs obtained by a multivariate permutation procedure with 100 random permutations of the assignment of the set of stage and grade. We selected the top 100 genes, so that the FDR was around 15%. Half of the selected genes were related by one of several functions such as transcription, proliferation, and immune system. Based on the estimates of regression coefficients, β_1 , β_2 , and β_3 , several patterns in differential gene expression can be figured out for the selected genes. We ignored the interaction if the estimate of β_3 was less than half of the maximum of the estimates of β_1 and β_2 in absolute value. Interestingly, 60 or more of the 100 selected genes had interaction. Approximately two-thirds of such genes had negative interactions. Figure 23.2 shows typical mean expression profiles for the selected genes with negative (Figure 23.2(a) and (c)) or positive interaction (Figure 23.2(b) and (d)). Negative (positive) interaction means under (over) expression for the tumors with the worst type, invasive and high grade, compared with an estimated level based on the main effects of stage and grade. Figures 23.2(a) and (b) may indicate differential expression between invasive tumors with high grade (the worst type) and the others, while Figures 23.2(c) and (d) may indicate differential expression between superficial tumors with low grade (the best type) and the others. A large part of the genes with interactions had mean expression profiles similar to Figures 23.2(a) and (b). Regarding the rest of the selected genes with no interaction, 10 or more genes had comparable size of effects for both β_1 and β_2 , and the rest had large effects only for β_1 or β_2 .

The selected genes have the potential to provide some clues from which genetic pathways in disease progression could be reconstructed. Different patterns in estimates of regression coefficients, β_1 , β_2 , and β_3 , would reflect different aspects in disease progression, and genes with similar patterns are suspected to an assemblage of genes coexpressed along pathways in disease progression. The selected genes will be further investigated in studies for pathway inference across a series of experimental conditions such as time.

The selected genes can be regarded as candidate markers that can potentially improve diagnostic accuracy of current approaches. For example, genes that had mean expression profiles such as Figures 23.2(a) and (b) have the potential to contribute to

TABLE 23.2
Estimates of FDR for Various P -Value Cut-Offs

P -value cut-off	Number of significant genes	Estimate of expected number of false positives	Estimate of FDR (%)
0.0500	942	302.0	32.1
0.0100	355	66.6	18.8
0.0050	221	35.0	15.8
0.0010	66	8.7	13.2
0.0005	38	4.8	12.6

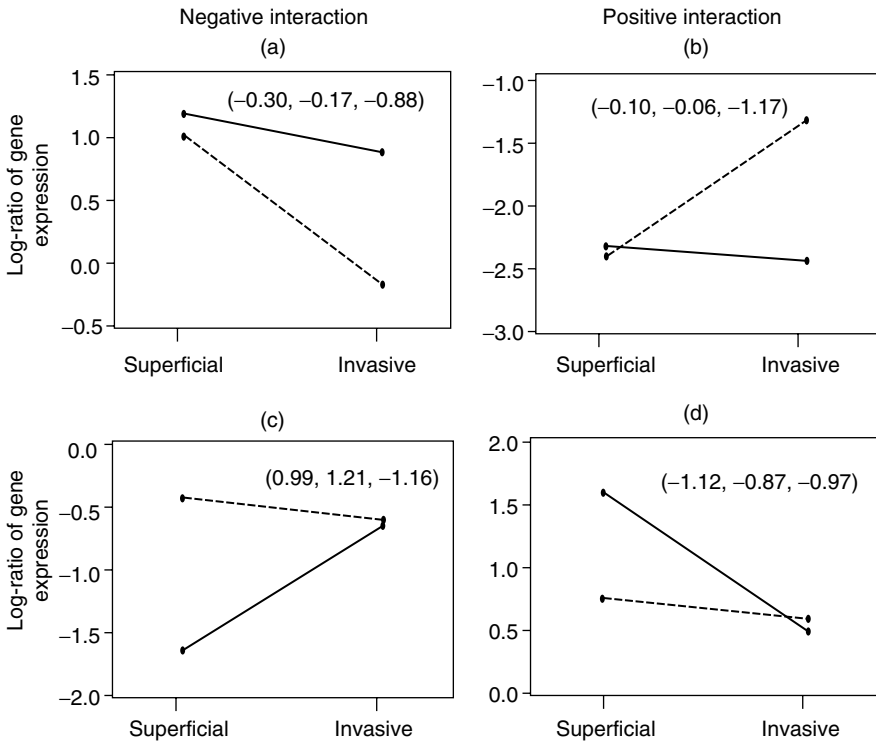


FIGURE 23.2 Typical expression profiles for genes with negative or positive interaction. Each plot shows the mean expressions of a gene of the 100 selected genes for the four types of tumors combined by stage, superficial and invasive; and grade, low (solid line) and high (dashed line). The estimates of parameters $(\beta_1, \beta_2, \beta_3)$ are also given.

classification of the invasive tumors with high grade (the worst type) and the others. We will develop expression-based predictors for this classification in Section 23.5. The selected genes will form the basis of this prediction.

23.5 PREDICTION

Another clinical application of DNA microarrays is the development of expression-based prediction systems that predict diagnosis classes or outcomes. A supervised approach is generally more powerful than an unsupervised approach. There are three important components in supervised prediction, selection of relevant genes for prediction, development of prediction models, and estimation of predictive accuracy.

23.5.1 SELECTION OF GENES

The simplest approach to select relevant genes is based on marginal association between each gene expression and the response variable as outlined in Section 23.4.^{6,7,9,13,14} More complex approaches include pair-methods,⁶⁶

application of genetic algorithms,^{67,68} and replication algorithms.⁶⁹ LeBlanc et al.⁷⁰ developed indices to link previously studied genes.

23.5.2 DEVELOPMENT OF PREDICTION MODELS

After filtering down to a subset of genes, prediction models can be formed. We outline a few common methods for class prediction that have been previously used including linear discriminant analysis (LDA),^{71,72} nearest-neighbors,^{71,73} classification trees,^{71,74} and support vector machines.^{75,76} See⁷⁷ for general discussions of class prediction methods. See^{78–82} for prediction of continuous survival outcomes.

Suppose one has a new sample with the vector of gene expression $x^{*'} = (x_1^*, \dots, x_G^*)$ for G selected genes and wishes to assign the sample to one of two classes, 1 and 2. The LDA method calculates

$$l(x^*) = (\bar{x}_1 - \bar{x}_2)' S^{-1} x^* \quad (23.2)$$

where $\bar{x}_i = (\bar{x}_{i,1}, \dots, \bar{x}_{i,G})'$ is the mean expression profile for the training set samples with class i for $i = 1, 2$ and S is an estimate of the pooled within-class covariance matrix of the expression levels of the genes. The new sample is predicted to be of the class that has the closer average of the linear combination in the training set to $l(x^*)$. In other words, the new sample is assigned to class 1 if $(\bar{x}_1 - \bar{x}_2)' S^{-1} \{x^* - (\bar{x}_1 + \bar{x}_2)/2\} > 0$ and to class 2 otherwise. A simpler form of LDA is diagonal linear discriminant analysis (DLDA), in which the diagonal matrix of S is used instead of S , resulting in the linear combination,

$$l_D(x^*) = \sum_{g=1}^G u_g x_g^* \quad (23.3)$$

where $u_g = (\bar{x}_{1,g} - \bar{x}_{2,g})/s_g^2$ for $g = 1, \dots, G$. Estimation of the many $G(G-1)/2$ parameters for correlations between gene expressions is not needed in DLDA. Some authors proposed to use variants of DLDA for prediction analysis of microarray data such as Golub's weighted vote method⁶ with a signal-to-noise statistic for u_g and the compound covariate predictors⁸³ with the standardized t -statistic for u_g . Prediction analysis of microarrays (PAM)⁸⁴ can also be regarded as a variant of DLDA.

The k -nearest-neighbors (k NN) approach is a simple nonparametric method that finds the k nearest neighbors in the training set closest to the gene expressions on a new sample and then decides the classification of this sample by majority vote.⁷⁷ The value of k is prespecified and typically chosen so that cross-validated predictive error is minimized.

Classification trees recursively portion the space of gene expression profiles into subsets that are each homogeneous with respect to levels of the response variable and form a flexible method to handle interactions between genes.^{77,85} A problem of single classification trees is that they tend to be unstable because of the hierarchical nature of the process: the effect of an error in the top split is propagated down to all of the splits below it.⁷⁷ *Aggregating* is an approach to improve the predictive capability in which slightly perturbed predictors from the training set are produced by

resampling and then aggregated to produce a composite predictor by voting.⁸⁶ *Bootstrap aggregating*, or *bagging*,⁸⁶ and *boosting*⁸⁷ are methods for generating perturbed versions of the training set.

Support vector machines (SVMs) are popular machine learning algorithms for binary classification.^{77,88} For the training set, SVMs construct a hyperplane for the decision surface as is done in LDA, but the hyperplane is determined such that the margin of separation between the classes is maximized, where the margin represents distance from the hyperplane to the closest correctly classified samples.

Analyzing a series of cancer datasets, Dudoit et al.⁷¹ reported that simple methods such as DLDA and *k*NN using a distance metric that does not account for correlation between genes performed well in term of predictive accuracy compared with more complex methods such as aggregated classification trees. Dudoit et al. argued that the poor performance of LDA is likely due to the poor estimation of correlations between genes with a small training set and a fairly large number of predictors.

23.5.3 ESTIMATION OF PREDICTIVE ACCURACY

Determination of predictive accuracy is particularly important when performing model selection in high dimensions. For class prediction problems, the proportion of correctly predicted or $1 - \text{misclassification proportion}$, sensitivity, and specificity are commonly used. A smoother index for the proportion of correctly predicted is the mean cross-validated log-likelihood.⁸⁴ For survival outcomes, the cross-validated log partial likelihood can be used.^{81,89}

Resampling techniques such as cross-validation and bootstrap⁹⁰ are very useful to assess predictive accuracy for the small sample sizes commonly encountered in microarray studies. When using these techniques, it is critical that all aspects of model building including selection of predictors are reperformed for each round in resampling.^{91,92} If selection of predictors and prediction models are optimized based on cross-validated predictive accuracy, an independent validation set is needed to have an unbiased estimate of the predictive accuracy because of the optimization process in model building for the training set. An independent validation set may also be needed when the model building process is complex and not easily specified in an algorithmic manner; see for example Rosenwald et al.¹⁴ and.⁹³

It is also important to establish that the predictive error is statistically lower than that expected when there is no relation between expression profile and the response variable. Radmacher et al.⁸³ proposed use of a permutation procedure to assess the statistical significance of a cross-validated predictive error.

23.5.4 ILLUSTRATION

We consider prediction of the binary class of whether patients have one of the worst tumors, invasive cancer (stage > T1) with grade of 3 (class 1) or not (class 0) based on 6,000 genes that passed a gene filtering step; see Section 23.2. Fifteen and thirty-three patients were involved in class 1 and class 0, respectively. Note that this class variable corresponds to the interaction term between stage and grade in the linear regression model (23.1) in Section 23.4. As a reminder, the interaction was

substantial for a large part of the top 100 genes. Thirty-nine genes were significant when the P -value cutoff $P=0.0001$ for a two sample t -test was used for all the 48 patients. Out of the 39 genes, 37 (95%) were also included in the top 100 genes selected by the linear regression analysis in Section 23.4.

We adopted the five predictors outlined in Section 23.5.2, LDA with (23.2), DLDA with (23.3), compound covariate predictor (CCP), 3-NN based on Euclidean distance, and linear SVM. In order to assess predictive accuracy, we performed leave-one-out cross-validation. For each training set with one sample left out, we performed gene filtering using the P -value cutoff $P=0.0001$ for a two sample t -test from scratch and fitted the prediction model. Then we used the fitted model to predict the class of the left-out sample and compared the predicted class with the observed class. Note that different genes or even different numbers of genes can be selected for each training set. For LDA, we limited the number of genes used in the model to 20 because using the full set of significant genes gave noninvertible covariance matrices, which is an inherent limitation of LDA. The analyses were performed using BRB ArrayTools (see Section 23.6).

The proportions correctly predicted for LDA, DLDA, CCP, 3-NN, and SVM were 66, 81, 79, 81, and 77%, respectively. DLDA and 3-NN performed the best, whereas LDA performed the worst. The results were parallel to those of a comparison study by Dudoit et al.⁷¹ They reported that the performance of LDA improved when the number of genes in the model was decreased. However, in our data, the proportion correctly predicted for LDA reduced to 52% when the limit in the number of genes was decreased to 10. We also applied a permutation procedure to assess the statistical significance of cross-validated predictive accuracy.⁸³ The P -values estimated by 2,000 permutations for LDA, DLDA, CCP, 3-NN, and linear SVM were 0.251, 0.027, 0.017, 0.023, and 0.046, respectively. The cross-validated predictive accuracy for LDA was not statistically significant at the significance level of 5%.

We also performed another more complex strategy for model building. First, we divided the 15(33) samples with class 1⁰ randomly into training and validation sets with 10(22) and 5(11) patients, respectively (the division ratio of 2:1). We fixed the number of genes to be used for prediction to be 5, 10, 30, 50, 100, or 300. For the training set with 32 patients, we chose the optimal number of genes on the basis of leave-one-out cross-validated predictive accuracy. We fitted a final prediction model with the chosen number of genes for the entire training set and then used it to predict the class of the validation set with 16 patients. We repeated the entire process with different random divisions into training and validation sets to assess the instability in the selected genes and the distribution of predictive accuracy.

Table 23.3 summarizes the results of 500 repetitions when DLDA was used. As expected, the optimized cross-validated predictive accuracy for the training set was optimistic. The difference between the optimized proportion of correctly predicted and the proportion from the validation sets across 500 repetitions was 9.3%. The range of the mean proportion of correctly predicted for the validation set across 500 repetitions was (25.0–93.8%). The variability was substantial (see also Figure 23.3), which indicates that the results from only a single division into training and validation sets would be unreliable, especially when the sample size is small. The proportion to which each gene is selected in the final model fitting may provide useful

TABLE 23.3
Frequency Distribution of the Number of Genes Chosen by Leave-One-Out Cross-Validation (LOOCV) for the Training Set and the Proportion Correctly Predicted from the Training and Validation Sets for 500 Repetitions of the Entire Processes of Model Building Using DLDA

The number of genes chosen by LOOCV	Frequency (%)	Proportion of correctly predicted (%)			
		The minimized leave-one-out cross-validated proportion		The proportion from the validation sets	
		Mean	Range	Mean	Range
5	141 (28%)	83.9	(68.8, 96.9)	72.3	(31.3, 87.5)
10	102 (20%)	83.9	(68.8, 93.8)	74.3	(31.3, 93.8)
30	92 (18%)	82.6	(68.8, 96.9)	73.2	(25.0, 93.8)
50	51 (10%)	81.1	(68.8, 90.6)	75.4	(56.3, 87.5)
100	49 (10%)	79.5	(62.5, 87.5)	75.8	(31.3, 93.8)
300	65 (13%)	82.5	(62.5, 93.8)	71.6	(25.0, 93.8)
Overall	500	82.8	(62.5, 96.9)	73.5	(25.0, 93.8)

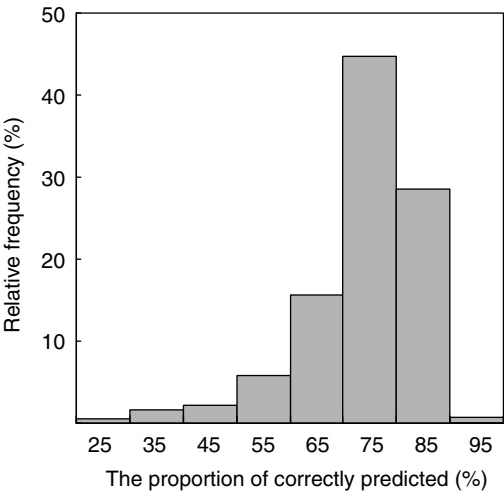


FIGURE 23.3 Frequency distribution of the proportion of correctly predicted from the validation set for 500 repetitions of the entire processes of model building using DLDA.

information for gene screening. The proportions for the top three genes with the highest proportions were 88%, 72%, and 53%. The proportions for the top 20, 50, and 100 genes were 29%, 19%, and 13%, respectively.

23.6 SOFTWARE NOTE

Software for commonly used unsupervised and supervised methods can be implemented in major statistical packages such as S-PLUS, R, and SAS. In S-PLUS and in free and open-source R packages, which can be downloaded from the Comprehensive

R Archive Network (<http://www.cran.r-project.org>), functions are available for specific methods. Functions for unsupervised clustering include hclust for hierarchical clustering, k -means for k -means clustering, pam for k -medioids, mclust for model-based clustering. Functions for supervised class prediction include lda for linear discriminant analysis, knn for k -nearest neighbor, svm for support vector machines. Classification trees can be implemented in the R package rpart.

Many free and open-source packages for analysis of microarray data have also been developed. EisenLab's Cluster is a popular package for hierarchical clustering, self-organizing maps, k -means, and principal component analysis, along with TreeView for displaying hierarchical clustering results using dendrograms (<http://rana.lbl.gov/EisenSoftware.htm>). GeneClust includes two basic functions for hierarchical clustering and gene shaving and simulation-based procedures to assess the results (<http://www.arraybook.org/>). SAM and PAM are software for multiple testing and class prediction, respectively (<http://www-stat.stanford.edu/~tibs/>). R Bioconductor provides useful packages for exploratory analysis and normalization, multiple testing and model-based analyses for gene identification, and class prediction (<http://www.bioconductor.org>). BRB ArrayTools is an integrated package for the visualization and statistical analysis of DNA microarray gene expression data, filtering, normalization, scatterplot, hierarchical clustering, multidimensional scaling, multiple testing, and class prediction (<http://linus.nci.nih.gov/~brb/>). See Parmigiani et al.⁹⁴ for other free software tools. For a more thorough list of software including commercial packages, see the Web site by Computer Science Department, The University of Dublin, <http://www.cs.tcd.ie/Nadia.Bolshakova/softwaretotal.html#P>.

ACKNOWLEDGMENTS

The author would thank Dr. Parry Guilford of Pacific Edge Biotechnology, Limited, New Zealand, and Professor Osamu Ogawa of Kyoto University Hospital for permission to use the data from the bladder cancer study. The author would also like to thank Drs. Hiroyuki Nishiyama, Jun Watanabe, and Masaaki Ito of Kyoto University Hospital and Dr. Seiichiro Yamamoto of the National Cancer Center Research Institute for many helpful discussions. Analyses in Sections 23.3 and 23.5 were performed using BRB ArrayTools developed by Dr. Richard Simon and Amy Peng.

REFERENCES

1. Ponder BA. Cancer genetics. *Nature* 2001;411:336–341.
2. Ramaswamy S, Golub TR. DNA microarrays in clinical oncology. *J Clinical Oncology* 2002;20:1932–1941.
3. Nguyen DV, Arpat AB, Wang N, Carroll RJ. DNA microarray experiments: biological and technical aspects. *Biometrics* 2002;58:701–717.
4. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000;403:503–511.

5. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 2000; 406:536–540.
6. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;286:531–536.
7. Pomeroy S, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, et al. Prediction of central nervous system embryonal tumor outcome based on gene expression. *Nature* 2002;415:436–442.
8. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Med* 2001;7:673–679.
9. Hedenfalk I, Duggan D, Chen Y, et al. Gene-expression profiles in hereditary breast cancer. *N Eng J Med* 2001;344:539–548.
10. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Nat Acad Sci USA* 2001;98:15149–15154.
11. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* 2000; 24:227–235.
12. Lakhani SR, Ashworth A. Microarray and histopathological analysis of tumors: the future and the past? *Nat Rev Cancer* 2001;1:151–157.
13. Wang E, Miller LD, Ohnmacht GA, Mocellin S, Perez-Diez A, Petersen D, et al. Prospective molecular profiling of melanoma metastases suggests classifiers of immune responsiveness. *Cancer Res* 2002;62:3581–3586.
14. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Eng J Med* 2002;346:1937–1947.
15. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RCT, et al. Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. *Nature Med* 2002;8:68–74.
16. Beer DG, Kardia SLR, Huang CC, Giordano TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Med* 2002;8:816–824.
17. Vasselli JR, Shih JH, Iyengar SR, Maranchie J, Riss J, Worrell R, et al. Predicting survival in patients with metastatic kidney cancer by gene-expression profiling in the primary tumor. *Proc Nat Acad Sci USA* 2003;100:6958–6963.
18. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *J Amer Stat Assoc* 1979;74:829–836.
19. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slides systematic variation. *Nucleic Acids Res* 2002;30:e15.
20. Gordon AD. *Classification*. New York: Chapman and Hall/CRC, 1999.
21. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Nat Acad Sci USA* 1998;95:14,863–14,868.
22. Yeung KY, Haynor DR, Ruzzo WL. Validating clustering for gene expression data. *Bioinf* 2001;17:309–318.
23. MacQueen J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967;1:281–297.

24. Kaufman L, Rousseeuw P. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley and Sons, 1990.
25. Kohonen T. *Self-Organizing Maps*. Berlin: Springer, 1997.
26. Tamayo P, Slonim D, Mesirov J, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Nat Acad Sci USA* 1999;96:2907–2912.
27. Brazma A, Vilo J. Gene expression data analysis. *FEBS Lett* 2000;480:17–24.
28. Getz GE, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. *Proc Nat Acad Sci USA* 2000;97:12,079–12,084.
29. Alon UN, Barkai DA, Notterman K, Gish S, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Nat Acad Sci USA* 1999; 96:6745–6750.
30. Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P. “Gene shaving” as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol* 2000;1:research0003.1–0003.21.
31. Lazzeroni L, Owen A. Plaid models for gene expression data. *Statistica Sinica* 2002; 12:61–86.
32. Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 1985;50:159–179.
33. Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol* 2002;3:1–36.
34. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *J Roy Stat Soc* 2001;B63:411–423.
35. Calinski R, Harabasz J. A dendrite method for cluster analysis. *Commun Stat* 1974; 3:1–27.
36. Krzanowski WJ, Lai YT. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics* 1988;44, 23–34.
37. Ben-Hur A, Elisseeff A, Guyon I. A stability-based method for discovering structure in clustered data. *Pacific Symp Biocomputing* 2002;7:6–17.
38. Yeung KY, Fraley C, Mutua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. *Bioinf* 2001;17:977–987.
39. McLachlan GJ, Bean RW, Peel D. A mixture model-based approach to clustering of microarray expression data. *Bioinf* 2002;18:413–422.
40. Pan W, Lin J, Le C. Model-based cluster analysis of microarray gene-expression data. *Genome Biol* 2002;3:research 0009.1–0009.8.
41. McShane LM, Radmacher MD, Freidlin B, Yu R, Li M, Simon R. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinf* 2002;18:1462–1469.
42. Kerr MK, Churchill GA. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc Nat Acad Sci USA* 2001; 98:8961–8965.
43. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Nat Acad Sci USA* 2000; 97:10,101–10,106.
44. Assersohn L, Gangi L, Zhao Y, Dowsett M, Simon R, Powles TJ, Liu ET. The feasibility of using fine needle aspiration from primary breast cancers for cDNA microarray analyses. *Clin Cancer Res* 2002;8:794–801.
45. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat* 1979; 6:65–70.
46. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988;75:800–802.
47. Westfall PH, Young SS. *Resampling-Based Multiple Testing*. New York: Wiley, 1993.

48. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stati Soc* 1995;B57:289–300.
49. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to ionizing radiation response. *Proc Nat Acad Sci USA* 2001;98:5116–5121.
50. Korn EL, Troendle JF, McShane LM, Simon, R. Controlling the number of false discoveries: application to high-dimensional genomic data. *J Stat Plan and Inference* (in press).
51. Storey JD, Tibshirani R. Estimating false discovery rates under dependence, with applications to DNA microarrays, Technical Report 28, Department of Statistics, Stanford University, Stanford, CA, 2001.
52. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* 2001;96:1151–1160.
53. Storey JD. A direct approach to false discovery rates. *J Roy Stat Soc* 2002; B64:479–498.
54. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. *J Roy Stat Soc* 2004;B66:187–205.
55. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinf* 2001; 17:509–519.
56. Newton MA, Kendzioriski CM, Richmond CS, Blattner FR, Tsui KW. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comput Biol* 2001;8:37–52.
57. Lonnstedt I, Speed TP. Replicated microarray data. *Statistica Sinica* 2001;12:31–46.
58. Broet P, Richardson S, Radvanyi F. Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *J Comput Biol* 2002; 9:671–683.
59. Ibrahim JG, Chen MH, Gray RJ. Bayesian models for gene expression with DNA microarray data. *J Am Stat Assoc* 2002;97:88–99.
60. Pan W, Lin J, Le CT. A mixture model approach to detecting differentially expressed genes with microarray data. *Funct Integr Genomics* 2003;3:117–124.
61. Wright G, Simon R. A random variance model for detection of differential gene expression in small microarray experiments. *Bioinf* 2003;19:2448–2455.
62. Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. *J Comput Biol* 2000;7:819–837.
63. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol* 2001;8:625–637.
64. Thomas JG, Olson JM, Tapscott SJ, Zhao LP. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res* 2001;11:1227–1236.
65. Kauermann G, Eilers P. Modeling microarray data using a threshold mixture model. *Biometrics* 2004;60:376–387.
66. Bo TH, Jonassen I. New feature subset selection procedures for classification of expression profiles. *Genome Biol* 2002;3:research 0017.1–0017.11.
67. Li L, Weinberg CR, Darden TA, Pedersen LG. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinf* 2001, 17:1131–1142.
68. Ooi, CH, Tan P. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinf* 2003;19:37–44.
69. Deutsch JM. Evolutionary algorithms for finding optimal gene sets in microarray prediction. *Bioinf* 2003;19:45–54.
70. LeBlanc M, Kooperberg C, Grogan TM, Miller TP. Directed indices for exploring gene expression data. *Bioinf* 2003;19:686–693.

71. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc* 2002;97:77–87.
72. Iizuka N, Oka M, Yamada-Okabe H, Nishida M, Maeda Y, Mori N, et al. Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *Lancet* 2003;361:923–929.
73. West M, Blanchette C, Dressman H. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Nat Acad Sci USA* 2001; 98:11,462–11,467.
74. Zhang H, Yu C-Y, Singer B, Xiong M. Recursive partitioning for tumor classification with gene expression microarray data. *Proc Nat Acad Sci USA* 2001;98:6730–6735.
75. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Nat Acad Sci USA* 2000;97:262–267.
76. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinf* 2000;16:906–914.
77. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer, 2001.
78. Nguyen DV, Rocke DM. Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinf* 2002;18:1625–1632.
79. Park PJ, Tian L, Kohane IS. Linking gene expression data with patient survival times using partial least squares. *Bioinf* 2002;18 Suppl 1:S120–127.
80. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. <http://www-stat.stanford.edu/~tibs/research.html>, 2003.
81. Pauler DK, Hardin J, Faulkner JR, LeBlanc M, Crowley JJ. Survival analysis with gene expression arrays. In *Handbook of Statistics*. Vol. 23. *Advances in Survival Analysis*. Amsterdam: Elsevier, 2004, 675–688.
82. Pawitan Y, Bjohle J, Wedren S, Humphreys K, Skoog L, Huang F, Amler L, Shaw P, Hall P, Bergh J. Gene expression profiling for prognosis using Cox regression. *Stat Med* 2004;23:1767–1780.
83. Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. *J Comput Biol* 2002;9:505–511.
84. Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Stat Sci* 2003;18:104–117.
85. Breiman L, Friedman J, Stone C, Olshen R. *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.
86. Breiman L. Bagging predictors. *Machine Learning* 1996;24:123–140.
87. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* 1997;55:119–139.
88. Vapnik V. *Statistical Learning Theory*. New York: Wiley, 1998.
89. Verweij M, Houwelingen V. Cross-validation in survival analysis. *Stat Med* 1993; 12:2305–2314.
90. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. New York: Chapman & Hall/CRC Press, 1993.
91. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Nat Acad Sci USA* 2002;99:6562–6566.
92. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Nat Cancer Inst*, 2003; 95:14–18.
93. Simon R, Korn EL, McShane LM, Radmacher MD, Wright G, Zhao Y. *Design and Analysis of DNA Microarray Investigations*. New York: Springer, 2004.
94. Parmigiani G, Garrett ES, Irizarry RA, Zeger SL. *The Analysis of Gene Expression Data: Methods and Software*. New York: Springer, 2003.

24 Profiling High-Dimensional Protein Expression Using MALDI-TOF: Mass Spectrometry for Biomarker Discovery

*Yutaka Yasui, Tim Randolph, and
Ziding Feng*

CONTENTS

24.1	Introduction	438
24.2	Overview of Cancer Biomarker Discovery via MALDI-TOF MS	438
24.2.1	MALDI-TOF MS Technology	438
24.2.1.1	Sensitivity	440
24.2.1.2	Resolution	440
24.2.1.3	Mass Accuracy and Precision	440
24.2.2	Study Design Considerations	440
24.3	Pre-Analysis Processing of MALDI-TOF Mass Spectra	441
24.3.1	Calibration of the m/z Axis	442
24.3.2	Peaks as the Principal Feature of Interest in Spectra	442
24.3.3	Imprecision in m/z Measurements	442
24.3.4	Adjusting for Baseline Intensity	444
24.3.5	Normalization/Standardization of Intensity	444
24.4	Evaluating Peaks Individually as Potential Markers	446
24.5	Forming a Biomarker Panel Using Multiple Peaks	447
24.5.1	Combination of Single-Peak Classification Rules	447
24.5.2	Combining Multiple Peaks into a Classifier	448
24.5.3	Building a Multi-Peak Single Classifier via Real AdaBoost	450
24.5.4	Extension to a Setting where Class Labels may Contain Error	451
24.6	Discussion	453
	References	454

24.1 INTRODUCTION

Biomarkers of cancer, such as carcinoembryonic antigen (CEA) for colorectal cancer and prostate-specific antigen (PSA) for prostate cancer, have a wide range of potential clinical utilities including early detection, disease monitoring, and prediction of treatment response. Discovery of novel tumor markers is an area of active research in oncology utilizing rapidly developing proteomic and genomic biotechnologies. Since proteins are closer to cell function than DNA and mRNA, it is logical to consider proteins (and peptides) as potential biomarkers. Measuring proteins is a complex task, however. First, proteins are dynamic. They are changed by interactions, posttranslational modifications, and perturbations from the environment. Thus, measurements are, strictly speaking, specific to cells, time, and environment. Second, in contrast to the digital nature of DNA and mRNA, proteins have three-dimensional structures and do not hybridize to their complements like DNA and mRNA. This makes the physical measurement of proteins and especially their quantification challenging. Third, proteins exist naturally in a mixture (e.g., in human serum), and some proteins (e.g., albumin and transferrin) dominate with respect to abundance in the mixture. This necessitates the separation of proteins in the mixture, especially if potential biomarkers are not as abundant as many other proteins in the mixture.

Many novel developments have been and continue to be made in proteomic technologies to overcome these complexities. An additional consideration for tumor marker discovery is the *throughput* of the technologies, broadly defined as the speed or rate at which the technologies can analyze samples. Estimating biomarkers' sensitivity and specificity requires an analysis of many samples covering a wide range of host and environmental conditions of cancer and cancer-free populations. High-throughput is, therefore, a necessary feature for a proteomic technology if it is to be applied to tumor marker discovery. Based on the throughput, some researchers chose the matrix assisted laser desorption and ionization (MALDI) time-of-flight (TOF) mass spectrometer (MS) as the tool for tumor marker discovery. This paper focuses on the study design and analysis methods of MALDI-TOF mass spectra in the context of tumor marker discovery. Note that we use the term MALDI-TOF MS broadly to include a popular TOF MS, surface enhanced laser desorption and ionization (SELDI)-TOF MS, which is a MALDI-TOF MS with samples mounted on a patented plate that is specially coated to bind proteins with specific properties (Ciphergen Biosystems, Inc.). While new biotechnologies are being developed rapidly, the principles discussed here apply to most other platforms of mass spectrometry.

24.2 OVERVIEW OF CANCER BIOMARKER DISCOVERY VIA MALDI-TOF MS

24.2.1 MALDI-TOF MS TECHNOLOGY

Mass spectrometers separate proteins/peptides from each other by their differences in mass-per-charge (m/z) ratio after ionization. MALDI-MS uses laser pulses to desorb biomolecules (proteins/peptides) to gas-phase ions with the aid of a "matrix" that is mixed with the biomolecules and absorbs/transfers the laser energy to ionize the

biomolecules without fragmenting. MALDI was developed by Tanaka and colleagues,¹ which led to a Nobel Prize in chemistry in 2002, using a metal powder matrix, and by Karas and Hillenkamp,² who proposed the use of an organic photoactive compound matrix, the matrix widely used today. The gas-phase ions are accelerated by a high-voltage supply between two electrodes based on their m/z ratio and then drift down a flight tube: a TOF-MS uses the differences in this transit time (time-of-flight), t , through the flight tube to separate ions of different masses. A detector placed at the end of the flight tube generates an electric signal when struck by an ion and measures the amount of hits at each t yielding the intensity values: ions with smaller masses reach the detector faster. The determination of m/z values in a TOF-MS is based on the following principle.

The magnitude of the force for an ion of charge z is zE where E is the electric field. This equals the kinetic energy of the ion, $mv^2/2$, where m is the mass and v is the flight speed of the ion, which gives rise to an equation:

$$z E = m v^2/2 = m (d/t)^2/2$$

where d is the flight distance and t is the flight time. Rewriting this equation gives a formula for m/z values:

$$m/z = 2 E(t/d)^2 = B (t - A)^2 \quad (24.1)$$

where A and B are calibration constants. The values of m/z in a TOF-MS are usually determined by estimating A and B in an internal or external calibration to known m/z values. It is these calibrated m/z values that are shown in the x -axis of TOF-MS output (see, for example, Figure 24.1).

Sample preparations of MALDI-TOF MS are outside of the scope of this monograph but are critical in the design considerations of MALDI-TOF-MS-based studies for tumor marker discovery. Important considerations in the sample preparations include removal of abundant proteins, protein separations, and choice and concentration of matrix. See Mann and Talbo³ for more information.

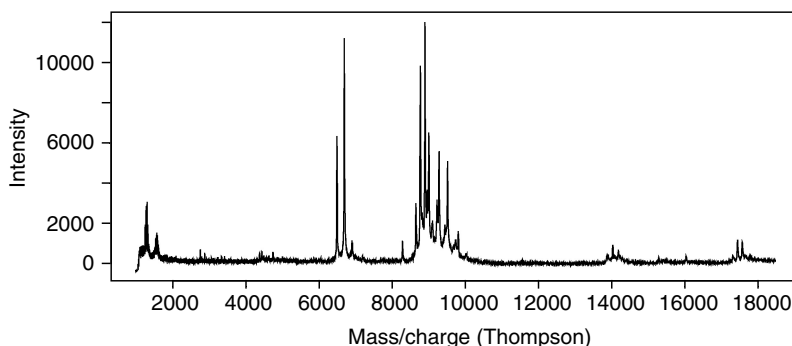


FIGURE 24.1 An example of MALDI-TOF mass spectra.

24.2.1.1 Sensitivity

Sensitivity refers to the ability to measure biomolecules with low concentrations. This is an important characteristic for MS employed in tumor marker discovery because markers may be present in very low levels in biological specimens. In fact, this is a major point of criticisms to MALDI-TOF-MS-based tumor marker discovery.⁴

24.2.1.2 Resolution

Resolution of an MS refers to its ability to distinguish different m/z values, commonly expressed as the peak intensity height divided by the width at the half height of the peak. Higher resolution means the ability to differentiate protein (peptides) with closer m/z values. Low resolution can be a problem if a tumor marker cannot be differentiated from proteins (peptides) with a similar mass, which can lead to attenuation of the marker's differential expression between cancer and cancer-free subjects.

24.2.1.3 Mass Accuracy and Precision

Mass accuracy refers to closeness between observed m/z values of a protein (peptide) and the true value. Unless we are using a MALDI-TOF MS for the purpose of protein identification, lack of mass accuracy itself is not a major concern: if the m/z of a protein is always measured to be the true value T plus 10 Thompson, we would recognize $(T + 10)$ Thompson as the protein's m/z value. Mass precision refers to variability of m/z measurements; that is, higher precision means a smaller variation of the m/z measurements around their average. Low precision of m/z measurements is a major concern in the analysis because a single protein (peptide) can appear with different m/z values and hence be labeled as a different protein (peptide) across mass spectra if m/z values are used as the labels of proteins (peptides).

24.2.2 STUDY DESIGN CONSIDERATIONS

Tumor marker discovery studies, at least in their early stages,⁵ use the “case-control” design where cases of certain cancer patients and controls are compared with respect to protein expression profiles. This seemingly simple study design has a number of potential methodological pitfalls that are highly relevant to tumor marker discovery studies. Epidemiologists have developed basic principles for case-controls studies to prevent such pitfalls; they can be summarized as follows:⁶ 1) a common study base for cases and controls; 2) controlling for confounding effects; and 3) comparable accuracy in “exposure” measurements (protein expression, in our case) between cases and controls.

The first principle requires defining a clear common study base (who, when, where) and sampling both cases and controls from it. Some tumor marker studies have used biological specimens of cases and controls from different sets of institutions. Others used archived case specimens and compared them with control specimens collected in a more recent screening program. These studies did not have a common study base, and therefore, cases and controls differ not only the cancer

status but also other factors (factors related to institutions and time) that could influence specimens and protein expressions in them.

The second principle requires controlling for confounding effects. Even if one uses a common study base, such as women age 40–70 currently residing in a defined community, cases of, say, breast cancer would tend to be older compared to the distribution of age in the rest of the women in the study base. If race/ethnicity is associated with the risk of cancer under study, racial/ethnic compositions of cases could differ from those of controls. In these examples, comparisons of protein expressions between cases and controls may be distorted by the imbalance of age and race/ethnicity because they are likely to influence protein expressions. Control for such confounding effects needs to be addressed either in the study design by stratified sampling of cases and controls, such as stratification by age or race/ethnicity to achieve a balance, or in the analysis stage by adjusting for these potential confounders, or both.

The third principle requires making measurement errors comparable between cases and controls. Study procedures including those in laboratories usually aim to minimize measurement errors. In case-control studies, it is critical that such procedures are applied nondifferentially to cases and controls. One must unify procedures for data and specimen collection, processing, storage, and assay between cases and controls, ideally blinding study staff and technicians with respect to the case-control status. If multiple MS machines, assay days, and technicians are used, their use must be balanced and perhaps randomized between cases and controls. These are efforts to minimize bias in the case-control comparison caused by differences in various steps of obtaining protein expression measurements. Without them, markers of any one of these factors could differentiate cases and controls, and the study could falsely regard them as potential tumor markers.

These remarks about the necessity of balance and randomization apply to any collection of case-control data, but the potential for bias is magnified here where a single datum is a spectrum consisting of tens of thousands of correlated measurements: each spectrum will contain substantial signal (that is, structure carrying information) that is not a manifestation of biological content or a product of tumor. The analyst must not blindly rely on an algorithm to distinguish between cancer-related signal and non-cancer-related signal.

24.3 PRE-ANALYSIS PROCESSING OF MALDI-TOF MASS SPECTRA

Figure 24.1 shows an example of MALDI-TOF mass spectra in the m/z range of 1,010–18,464 Thompson. There are 31,442 measurement points (m/z , intensity) in this m/z range. These data have complex features due to complexities in both biological specimens and interfering biochemical/physical processes of the measurement procedure. In extracting relevant information from such high-dimensional complex data, one must recognize some key issues/properties associated with MALDI-TOF mass spectra described below. See Baggerly et al.⁷ and Sorace and Zhan⁸ for informative discussions of various potential issues in the application of MALDI-TOF MS for tumor marker discovery.

24.3.1 CALIBRATION OF THE m/z AXIS

First, care must be exercised in the calibration of the m/z axis. Typically, the calibration involves measuring a standard sample that contains several species with known m/z values and fitting the quadratic curves of Equation (24.1) to the observed versus known m/z values of the several species to estimate the constants A and B. If the fit of the quadratic curve is very good, it is an indication that the calibration method worked. If the fit is rather poor, however, a careful examination is needed. For example, the fit could be improved appreciably by removing one poorly measured species from the calibration procedure. We have encountered such a situation where we removed the problematic species from the calibration and achieved a clear improvement in the quality of subsequent analyses.

A philosophically different approach is to use the raw TOF measurements instead of the calibrated m/z values; see, for example, Baggerly et al.⁹ Because the calibration imposes an artificial construct in terms of the quadratic model and the fitting of the model between the actual TOF measurements and the analysis, it may be advantageous to forego calibration. However, the imprecision problem of the m/z values discussed below also exists in raw TOF measurements, and a proper handling of this problem needs to be addressed.

24.3.2 PEAKS AS THE PRINCIPAL FEATURE OF INTEREST IN SPECTRA

Peaks in MALDI-TOF mass spectra correspond to proteins or peptides with measurable quantities. They present features that are biologically interpretable. This notion led some investigators to define peaks mathematically from spectra and focus subsequent analyses on them, while others analyzed spectra data without defining peaks. Yasui et al.^{10,11} and Tibshirani et al.¹² used a simple method of identifying peaks by searching local maxima in spectra. Coombes et al.¹³ defined peaks elegantly using both local maxima and local minima addressing concerns regarding noise in spectra. Randolph and Yasui¹⁴ took a different view and defined features in terms of local changes that occur at various scales. As defined there, the features are not restricted solely to peaks. Indeed, the term *peak* is somewhat ambiguous when applied to MALDI-TOF mass spectra; because of the high degree of irregularity, local maxima occur in nearly any width of window at any location along a spectrum. Therefore, one of the goals of Randolph and Yasui¹⁴ was to provide a way of unambiguously defining and locating features of certain types. These features include scale-based local maxima, or peaks, loosely speaking, and shoulders that may not appear as local maxima. This was done using the notion of modulus maxima in a multiscale wavelet decomposition of each spectrum; see, for example, Mallat.¹⁵ The definition is not claimed to be correct or optimal but emphasizes that processing of spectra could proceed using a well-defined, flexible concept of features.

24.3.3 IMPRECISION IN m/z MEASUREMENTS

Even after a very good calibration of the m/z axis, observed m/z values of a given protein or peptide are not always recorded exactly the same across multiple measurements. In our quality-control experiments, observed m/z values fluctuate from

experiment to experiment by approximately ± 0.1 – 0.2% of the true m/z value depending on the quality of the MALDI-TOF MS machines and experimental conditions. Note that this is not solely a uniform shifting of the m/z axis; rather, the variation is more stochastic in nature. Note that this problem also exists in raw TOF measurements before calibration. An important consequence of this property is that individual proteins cannot be indexed properly by their observed m/z values for recognition across multiple spectra. Figure 24.2 illustrates this key property of MALDI-TOF mass spectra. Figures 24.2(a) and 24.2(b) show the same set of five MALDI-TOF mass spectra of an identical serum sample. Figure 24.2(b) is an enlargement of Figure 24.2(a) around the m/z value of 6,700 Thompson. The sample was spotted on five different wells of a plate, and each spotted well was measured by MALDI-TOF MS. While the five spectra in Figure 24.2(a) show common global features and peaks, a careful examination of a peak around 6,700 Thompson in

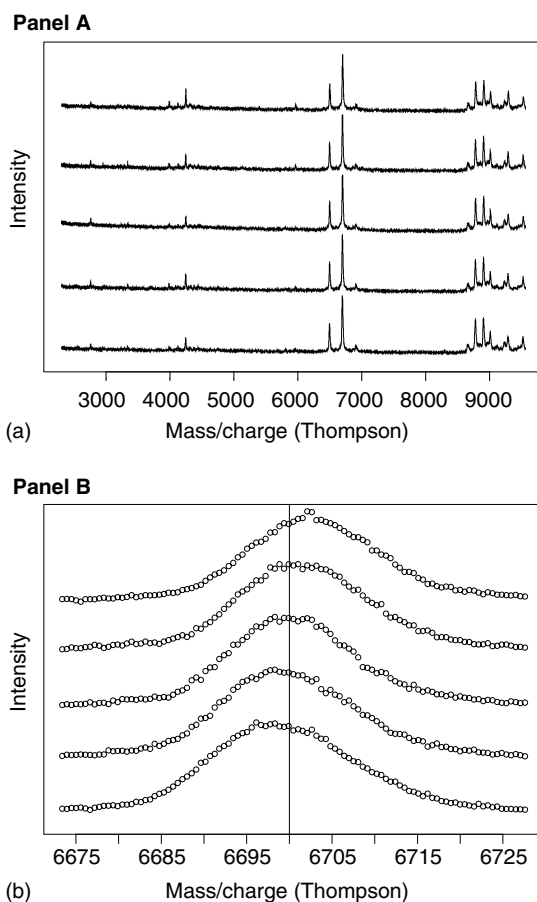


FIGURE 24.2 Five spectra of an identical serum sample: (a) global agreement of peaks across the five spectra; (b) a magnified version of (a) around 6,700 Thompson shows a small amount of shifting along the mass/charge axis.

Figure 24.2(b) reveals a small amount of local shifting of the peak along the m/z axis, about ± 10 points of measurements, illustrating the inability of observed m/z , or TOF values, to index the peak consistently.

To overcome the imprecision problem in the analysis stage, one must employ a procedure for recognizing and aligning different but close m/z values as the same species in an appropriate manner, which is closely related to the notion of defining peaks. The following authors proposed methods to handle this problem. Yasui et al.¹¹ proposed a simple method for aligning peaks across spectra with an m/z -window of a fixed width that corresponds to the fluctuation amount of m/z values (usually ± 0.1 – 0.2% of m/z). Tibshirani et al.¹² aligned peaks across spectra by a clever application of complete-linkage hierarchical clustering. Randolph and Yasui¹⁴ also addressed the problem of feature misalignment across spectra by using the distribution of feature locations in the entire data set; a histogram of feature locations provided a spectral-density-like function from which regions of single peptide masses were inferred. The method is illustrated in Figure 24.3. Figure 24.3(a) exhibits MALDI-TOF spectra taken from four different human serum samples in a study. The tickmarks at the bottom of the panel indicate the locations of their scale-5 features as defined in Randolph and Yasui.¹⁴ Figure 24.3(b) shows a histogram of scale-5 feature locations from the collection of all 220 spectra in this study. The dark curve is one type of local average of the histogram; its local minima were used to obtain the tickmarks at the bottom of Figure 24.3(b), which delimit scale-5 bins. Features from different spectra that are located within the same bin are treated as coming from the same protein. A feature of this approach is that it does not require prespecification of a fixed degree of imprecision in the m/z values.

24.3.4 ADJUSTING FOR BASELINE INTENSITY

The baseline intensity refers to the amount of elevation in intensity measurements caused by the matrix comprising pure small energy-absorbing molecules used in MALDI. The matrix elevates intensity measurements particularly in the low m/z region. Most MALDI-TOF mass spectrometers have an implemented baseline-subtraction algorithm that is functional. To use the intensity measurements quantitatively, the baseline intensity must be subtracted by the machine's internal algorithm or an external algorithm specifically developed for this purpose. See Coombes et al.¹³ for a simultaneous method for peak identification and baseline subtraction. An alternative approach, as offered in Randolph and Yasui,¹⁴ is to measure local changes rather than raw intensities. Defining a peak or feature as a local change in intensity means that the baseline intensity does not explicitly enter into the quantification of peaks.

24.3.5 NORMALIZATION/STANDARDIZATION OF INTENSITY

There are substantial variations in the absolute values of intensity measurements. It is generally agreed that spectrum intensity measures relative, not absolute, quantity of proteins because the extent of ionization is protein-specific and is affected by competition from neighboring proteins. In a simple quality-control experiment in which the samples are fixed to contain only a few proteins, coefficients of variations in the repeated intensity measurements of these proteins were approximately halved

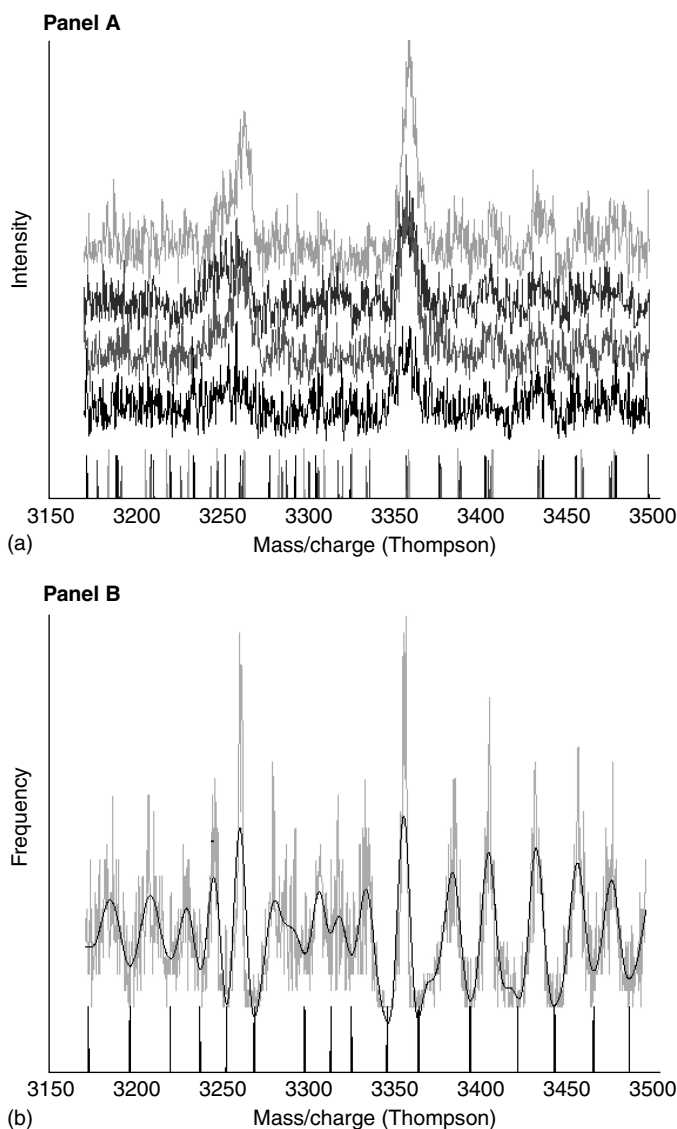


FIGURE 24.3 The peak alignment method of Randolph and Yasui (14).

if relative values of intensity, obtained by intensity measures divided by an intensity measure at a given mass, rather than absolute values of intensity were considered. Thus, it seems advantageous to normalize intensity measurements before making comparisons across spectra. Intensity measurements in a mass spectrum are often normalized at the time of measurement (that is, before the baseline subtraction) by the total ion currency of the spectrum, the sum of the separate ion currents carried by ions contributing to the spectrum. This converts absolute intensity measurements

to relative values. Further normalization/standardization at the analysis stage, possibly after the baseline subtraction, may be useful. Research is needed to determine appropriate methods. Logarithmic transformation is often used to achieve a less-skewed distribution for intensity measures.

24.4 EVALUATING PEAKS INDIVIDUALLY AS POTENTIAL MARKERS

In our view, the preanalysis processing of mass spectra is a step that is completely independent of the case-control status of samples. That is, each spectrum is processed equally through a unified preanalysis procedure regardless of whether it is from a case sample or from a control sample. Following the preanalysis processing, peaks that have been aligned can be evaluated for their discriminatory power for cases versus controls. The preanalysis processed dataset takes the form $\{y_{ij}; i = 1, 2, \dots, N, j = 1, 2, \dots, K\}$ where y_{ij} denotes the intensity value after the preanalysis processing for the i th spectrum's j th aligned peak. The N spectra may not be independent samples depending on the study design. Some may represent replicate measures of a single sample or pairs taken from each patient before and after treatment. Such potential dependences must be accounted for in certain study designs. In the assessment of the case-versus-control design, discriminatory power for the j th individual peak is essentially a two-class classification using $\{y_{ij}; i \in \text{CASE}\}$ and $\{y_{ij}; i \in \text{CONTROL}\}$ where CASE and CONTROL denote the index sets of cases and controls, respectively. In the independent case, the discriminatory power can be represented by the area under the receiver operating characteristic curve (AUC), which is equivalent to the Wilcoxon-Mann-Whitney U statistic:¹⁶

$$U_j = \frac{\sum_{k \in \text{CASE}} \sum_{l \in \text{CONTROL}} I\{y_{kj} > y_{lj}\}}{\text{Total number of (CASE, CONTROL) pairs}}$$

where $I\{a > b\}$ is an indicator function of the condition $a > b$. By the use of AUC or other appropriate measures, one can order the K peaks in terms of the case-control discriminatory power. Such descriptions of discriminatory power for *all* peaks are useful for comparisons across studies. Given the high dimensionality and complexity of the MALDI-TOF data, reporting only the best-performing peak(s) in each study would not permit effective and thorough comparisons of discriminatory power across studies.

Given the lack of an established normalization/standardization procedure, the peak probability contrast (PPC) method proposed by Tibshirani et al.¹² presents a practical measure of discriminatory power with a proper handling of highly variable intensity measurements. After identifying peaks and aligning them via a clustering algorithm, the PPC computes at each cluster of peaks j the proportion of cases $P_{aj}(\alpha)$ and controls $P_{oj}(\alpha)$ with intensity values exceeding various quantiles α . It then selects the quantile that maximizes the case-control difference of this proportion: $|P_{aj} - P_{oj}|$. This quantity can be used as the measure of discriminatory power instead of the AUC. Note that the PPC uses quantiles, while the AUC uses values, of the intensity after the preanalysis processing.

One can also statistically explore whether the discriminatory power of the j th peak is above a certain level. A significance test for the AUC of a given peak can be performed by the method of DeLong et al.¹⁷ or the bootstrap. If each sample was measured multiple times by MALDI-TOF MS, the bootstrap could be applied to clusters of such replicates assuming independence across clusters.

A critical issue in performing hypothesis tests for multiple peaks is the problem of multiple comparisons (see Hsu et al. and Matsui in this volume). False discovery rates (FDRs) proposed by Benjamini and Hochberg¹⁸ provide a framework that is different from the classic description of the probability of making one or more false discoveries. In our case, a discovery is a claim for a peak to have discriminatory power above a certain level. The FDR framework describes the expected proportion of false discoveries among all claimed discoveries. For a specific dataset, FDR can be estimated by a permutation procedure.¹² Specifically, an estimate of FDR can be obtained by dividing the expected number of false discovery by the number of discoveries in the original dataset. The expected number of false discoveries is given by averaging the number of discoveries in random permutations of case-control labels. Storey¹⁹ introduced the concept of q -value that assigns a significance level to each test in the setting of multiple testing. A q -value is a conservative estimate of the probability for a given claim of discovery to be false.

24.5 FORMING A BIOMARKER PANEL USING MULTIPLE PEAKS

A single peak used as a tumor biomarker has limited power. This is because there are multiple pathways to cancer even for a specific tumor type, genetic variations among individuals, and redundancy/compensatory mechanisms in biological systems. There are two conceptually different approaches to using multiple peaks as a biomarker panel. One is to combine multiple single-peak classification rules as one classification rule. The other is to form one classifier using multiple peaks with associated free parameters.

24.5.1 COMBINATION OF SINGLE-PEAK CLASSIFICATION RULES

Suppose there are M peaks following the preanalysis processing of MALDI-TOF mass spectra. Each peak could serve as an individual biomarker using its intensity and appropriate cutoff either below or above for predicting cancer cases, although its classification performance may be poor. A composite rule that requires meeting a set of multiple single-peak classification rules for cancer simultaneously (combining single-peak rules with *and*) would increase the specificity while decreasing the sensitivity, of each of the single-peak classifications. A composite rule that requires meeting one in a set of multiple single-peak classification rules for cancer (combining rules with *or*) would increase the sensitivity while decreasing the specificity of each of the single-peak classifications. Thus, one potential strategy for building a panel of biomarkers for a specific type of cancer is:

- Step 1, Combine multiple single-peak rules using *and* to target a subtype of the cancer type of interest with very high specificity; and
- Step 2, Combine multiple subtype-rules from Step 1 using *or* to cover multiple subtypes of the cancer type of interest without appreciably losing specificity.

24.5.2 COMBINING MULTIPLE PEAKS INTO A CLASSIFIER

Instead of combining single-peak classification rules into one rule, one may also consider forming a classification rule using multiple peaks with associated free parameters to be estimated. Figure 24.4 illustrates this concept. Neither peak A nor peak B can separate cancer cases and controls completely using its intensity values. That is, for both peaks, the marginal distributions of intensity for cases and controls are overlapping. However, in a linear combination of the intensity values for peak A and peak B, cases and controls are completely separable. Specifically, the joint distributions of peak-A intensity and peak-B intensity are not overlapping for cases and controls. In Figure 24.4, the multipeak single classifier is a straight line, α peak-A intensity + β peak-B intensity, and a cutoff value c gives a classification rule α peak-A intensity + β peak-B intensity $> c$. The free parameters, α and β , are estimated using a certain optimality criterion.

Classification performance of multipeak single classifiers could appear falsely high in a given dataset because overfitting occurs if they contain too many free parameters relative to the information in the dataset. This is particularly so when the dimension of MALDI-TOF mass spectra remains high even after the preanalysis processing. An overfit occurs for an excessively flexible classifier because it could adapt to any observed case-control differences regardless of whether they are

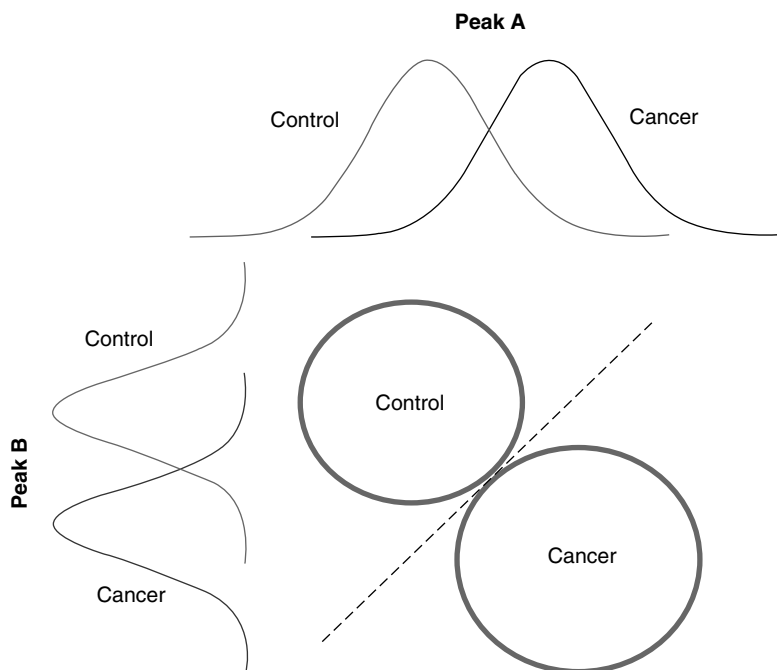


FIGURE 24.4 An advantage of using multiple-peak single classifier over each of its single components: the marginal distributions of markers A and B cannot separate cases and controls, while their joint distribution can.

consistently observed true differences or noise. Thus, a higher level of flexibility gained by using more peaks and associated free parameters has a trade-off of becoming more vulnerable to the overfitting problem.

There are two issues to consider in handling overfitting. One is how to build a multippeak single classifier keeping the chance of overfitting low. The other is how to measure classification performance without bias and with high precision. Three common strategies for handling overfitting consider the two issues differently. The first, and the simplest, strategy is to restrict the consideration to a family of multippeak single classifiers that has a limited degree of flexibility and hence low chance of overfitting. This strategy focuses on the first issue. The second strategy, emphasizing the second issue, ensures unbiased evaluation of classification performance by cross-validation or other techniques and uses the unbiased evaluation in selecting multippeak single classifiers with high classification power from a pool of candidates. The third strategy minimizes an objective function that reflects the trade-off between flexibility and overfitting risk by including both a term for goodness-of-fit and a penalty term for the number of parameters relative to the information in the dataset. A sequence of multippeak single classifiers with varying degrees of flexibility are considered, and the one that minimizes the objective function is selected.

Examples of the first strategy include Baggerly et al.⁹ and Yasui et al.¹⁰ Specifically, Baggerly et al. restricted the number of peaks to five or less in the family of multippeak classifiers in a protein biomarker discovery study for lung cancer. Yasui et al. used Real AdaBoost,²⁰ an algorithm known to resist overfitting by slowly modifying the classifier iteratively in a restricted fashion, in a protein biomarker discovery study for prostate cancer. See more on Real AdaBoost below.

Ruczinski et al.²¹ proposed a classifier building method called logic regression, which uses the second strategy. Specifically, their method employs k -fold cross-validation or an independent test set to measure classification performance without bias and to select the best size (flexibility) of logic-regression classifiers for a given dataset. Another example of the second strategy was in Tibshirani et al.'s PPC method¹² described earlier. Recall that the PPC computes at each cluster of peaks j the proportion of cases $P_{aj}(\alpha)$ and that of controls $P_{oj}(\alpha)$ with intensity values exceeding various quantiles (α s) and selects the quantile that maximizes the case-control difference of this proportion $|P_{aj} - P_{oj}|$. These group-specific proportions exceeding certain intensity thresholds form centroid vectors of the two groups, $(P_{a1}, P_{a2}, \dots, P_{aJ})$ and $(P_{o1}, P_{o2}, \dots, P_{oJ})$, where J is the number of peaks. A new spectrum can be classified by the nearest centroid classification rule. For feature selection and improvement of classification, the PPC uses soft-thresholding of the group proportions. It is this threshold that is chosen by an unbiased estimation of classifier performance using a 10-fold cross validation.

An example of the third strategy can be seen in a popular machine learning principle, structural risk minimization proposed by Vapnik.²² The approach is based on the following upper bound of expected classification error (risk) that involves a quantity h called Vapnik-Chervonenkis (VC) dimension:

$$\text{Expected error} \leq \text{observed error} + \sqrt{\frac{h\{\log(2n/h) + 1\} - \log(\eta/4)}{n}}$$

where the inequality holds with probability $1 - \eta$ and n is the number of observations. The principle of structural risk minimization considers a sequence of classifiers with varying VC dimensions, h , (i.e., flexibility) and selects the one that minimizes the upper bound for a given small η . See Burges²³ for an excellent tutorial on this principle and its link to support vector machines. Note that while the upper bound of the risk may be very conservative the principle may serve as a useful classifier selection strategy given the VC dimensions: the calculation of VC dimensions are often difficult, however. For the definition and calculation of VC dimensions, see Burges.²³

Regardless of how multippeak single classifiers are constructed using a training dataset, to estimate the final classifier's performance without bias it is useful to have a separate test dataset that was not used in any part of the construction of the final classifier. The test set must be saved in a lock box until final classifiers to be tested are completely specified using a training set. For example, using a test set for evaluating multiple classifiers that have been developed on a training set and reporting the ones with the best test-set performances would lead to an overestimation of the true performances of the reported classifiers because the test set was used in the selection of the reported classifiers. Given a single dataset, a training set and a test set can be constructed by random sampling stratified by the case-control status.

24.5.3 BUILDING A MULTI-PEAK SINGLE CLASSIFIER VIA REAL ADABOOST

We describe an application of the Real AdaBoost algorithm²⁰ to the construction of multippeak single classifiers (Yasui et al.).¹⁰ The boosting algorithm constructs a powerful summary classifier S by taking a majority-vote of many weak base classifiers, $C_1, C_2, C_3, \dots, C_M$ ($S_M = C_1 + C_2 + C_3 + \dots + C_M$). This is accomplished in a restricted stage-wise selection of base classifiers. Specifically, it performs a weighted selection of a base classifier at each stage given all the previously selected base classifiers and their associated parameters fixed. In selecting the next base classifier, C_m , at the m th stage, higher weights are given to observations that are incorrectly classified by the current summary classifier S_{m-1} , so that the new base classifier C_m will have a tendency toward correctly classifying the previously incorrectly classified observations. The utility of the boosting algorithm has been shown in a wide range of classification problems; see Friedman et al.²⁰ and Hastie et al.²⁴ for insightful discussions of the boosting algorithm. In Yasui et al.,¹⁰ the Real AdaBoost was applied to the construction of a summary multippeak single classifier

S_M of the following form: $S_M = \sum_{j=1}^M C_i = \sum_{j=1}^M (\alpha_j + \beta_j X_j)$, where the α s and β s are free parameters and X s are peak intensities (after the preanalysis processing) at M selected peaks, $\{m_j; j = 1, \dots, M\}$. The construction of S_M proceeds as follows. Let $Y = 1$ and $Y = -1$ denote cases and controls, respectively, and let N denote the total number of observations available in the training dataset. The algorithm starts with an equal weight for each observation, $W_{1i} = 1/N$, where i is the index for observations. In the first stage, the most statistically significant peak, m_1 , is selected by a likelihood-ratio test for a linear logistic regression model for the binary outcome Y with weights $\{W_{1i}\}$. The linear predictor of the logistic regression model forms the

best base classifier $S_{1i} = C_{1i} = \alpha_1 + \beta_1 X_{1i}$. In the second stage, the weight for each observation is updated by $W_{2i} = W_{1i} \times \exp\{-y_i S_{1i}/2\}$ and renormalized such that

$\sum_{j=1}^N W_{2i} = 1$. A weighted logistic regression with updated weights $\{W_{2i}\}$ chooses the

most significant peak m_2 by a likelihood-ratio test. An important feature of the algorithm is that the intensity X_1 of peak m_1 selected in the first-stage does not enter into the second-stage logistic regression as an explanatory variable. The only influence of X_1 on the second-stage selection is through the fixed weights $\{W_{2i}\}$. The summary classifier resulting from the second stage is $S_2 = C_1 + C_2$. This iterative process is repeated without any prespecified limit. Friedman et al.²⁰ viewed the above algorithm as a stage-wise minimization of a weighted negative Bernoulli log-likelihood. That is, at the m th stage, select a peak and associated parameters by:

$$(\alpha_m, \beta_m, X_m) = \underset{(\alpha, \beta, X)}{\operatorname{argmin}} \sum_{i=1}^N \exp \left\{ -y_i \sum_{j=1}^{m-1} (\alpha_j + \beta_j X_{ji})/2 \right\} \\ \times \log[1 + \exp\{-Y_i(\alpha + \beta(X)_i)\}]$$

where X_{ji} denotes the intensity value of the j th peak for the i th observation. A 10-fold cross-validation of the training set can be used to assess the classification performance of the m th stage classifier. Selecting a stage M with the best cross-validated classification performance will give the final classifier for which an unbiased estimate of classification performance is given by subjecting it to the test set. The cross-validated performance of the best classifier in the training set is not unbiased since the cross-validation was used for the selection of the best classifier. Therefore, a test set that is used only for the assessment of classification performance is necessary.

24.5.4 EXTENSION TO A SETTING WHERE CLASS LABELS MAY CONTAIN ERROR

In tumor marker discovery, errors in class labels are common. For example, biopsies for confirming cancer may miss cancerous cells. Therefore, cancer cases, especially those in early stage, can be labeled incorrectly as controls. In prostate cancer, this is known to occur.²⁵ one discover biomarkers for the correct labels? Here we describe briefly an extension of the boosting algorithm above to this mislabeled setting.²⁶ The boosting algorithm above can be rewritten as:

$$\phi_m = \underset{\phi}{\operatorname{argmin}} \sum_{i=1}^N f(Y_i, \phi; \theta_{m-1}) \quad (24.2)$$

where $\phi_m = (\alpha_m, \beta_m, X_m)$ and $\theta_m = (\phi_1, \phi_2, \dots, \phi_m)$. Recall that θ_{m-1} is fixed from the previous stage. Yasui et al.'s proposal was to treat the correct label Y_i^* of the i th observation as missing if its observed label Y_i is potentially incorrect and apply an EM-like algorithm to the observed incomplete data. If the diagnostic method used for labeling of cancer cases could miss some cancer cases, the proposal would treat the correct label Y_i^* of all controls as missing. When Y_i^* s are observable

(i.e., $Y_i^* = Y_i$ for all observations), we use (24.2) minimizing $\sum_{i=1}^N f(Y_i^*, \phi; \theta_{m-1})$. In the potentially mislabeled setting (i.e., $Y_i^* \neq Y_i$ for some observations), an expectation of $\sum_{i=1}^N f(Y_i^*, \phi; \theta_{m-1})$ is minimized, conditional on the observed labels $\{Y_i\}$ and the current parameter estimates from the previous stage. Knowledge about the nature of potential mislabeling could specify the set of observations with $Y_i^* = Y_i$, for which the conditional expectation of $f(Y_i^*, \phi; \theta_{m-1})$ given the observed label is equal to $f(Y_i, \phi; \theta_{m-1})$. For example, if the diagnostic method used for labeling of cancer cases is 100% specific (i.e., no false positives), all cancer cases are known to have correct labels, i.e., $Y_i^* = Y_i$. The modified algorithm, EM-Boosting, is:

$$\phi_m = \underset{\phi}{\operatorname{argmin}} \sum_{i=1}^N \sum_{Y_i^*=-1}^{Y_i^*=1} P(Y_i^* | Y_i, \theta_{m-1}) f(Y_i^*, \phi; \theta_{m-1})$$

The basic idea is to use a weighted loss function with weights being the current best estimates of case-control status given the observed data and parameter estimates; see Yasui et al.²⁶ for more details.

To illustrate the utility of EM-Boosting, the following simulation study was conducted using real MALDI-TOF MS data. The original dataset consisted of 245 disease cases and 81 disease-free controls, where the potential for mislabeling is known to be negligible. To mimic a real setting where mislabeling could occur by missing cancerous cells by biopsy-based diagnosis (i.e., mislabeling could occur as false negatives but not as false positives), we created two artificial mislabeled datasets by mislabeling a simple random sample from the disease group as disease-free. Two sizes of mislabeled samples were considered: 20% ($N = 49$) and 40% ($N = 98$) of the 245 disease cases. Thus, the first dataset had 196 ($= 245 - 49$) observations labeled correctly as disease cases and 130 ($= 81 + 49$) observations labeled as disease-free, of which 49 (37.7%) were mislabeled disease cases. Similarly, the second dataset had 147 ($= 245 - 98$) observations labeled correctly as disease cases and 179 ($= 81 + 98$) observations labeled as disease-free, of which 98 (54.7%) were mislabeled disease cases. EM-Boost was applied to the data with the observed labels without knowing the fraction of mislabeling. The performance was compared with two methods that ignore the mislabeling: logistic regression with the forward-variable selection with Bayesian information criterion (BIC) as the stopping rule and Real AdaBoost (24.2). The performance was evaluated in the artificially mislabeled datasets first against the true disease labels and then assessed in an independent test dataset of 45 disease cases and 15 disease-free subjects.

The results shown in Table 24.1 illustrate the utility of EM-Boost. Specifically, the EM-Boost classifiers had significantly larger AUCs and considerably higher sensitivity at 95% specificity than the forward-selection BIC classifier and the simple Real AdaBoost classifier. In the mislabeled data 1 where 37.7% of nondisease labels were incorrect, the EM-Boost classifier recovered over 95% of correct labels in the training set and was able to correctly classify approximately 90% of the samples in the test set. Even in the mislabeled data 2, where the proportion of incorrect labels

TABLE 24.1

Results of a Simulation Study Comparing EM-Boost vs. Real Adaboost and Logistic Regression with Forward-Selection Stopped by Bayesian Information Criterion

	Forward Selection with BIC	Boosting	EM-Boost***
Misabeled data 1*			
Training set			
AUC	0.958	0.974	0.992–0.993
Sensitivity at 95% specificity	64.5%	79.0%	96.3–97.5%
Test set (N = 60)			
AUC	0.807	0.816	0.919–0.936
Classification error	19/60	15/60	5–7/60
Misabeled data 2**			
Training set			
AUC	0.906	0.946	0.962–0.981
Sensitivity at 95% specificity	50.6%	58.0%	75.3–82.7%
Test set (N = 60)			
AUC	0.671	0.719	0.880–0.920
Classification error	28/60	26/60	8–12/60

* In Misabeled Data 1, 37.7% of nondisease labels were incorrect.

** In Misabeled Data 2, 54.7% of nondisease labels were incorrect.

*** EM-Boost was started with an initial value of the incorrect label fraction at 10, 30, and 50% in the nondisease label group. The ranges of the three results are shown.

in the nondisease group exceeded 50%, the EM-Boost classifier was able to classify 80% or more samples in the test set correctly.

24.6 DISCUSSION

We discussed proteomic-biomarker discovery analyses using data from MALDI-TOF MS. Three main steps of such analyses are preanalysis processing of mass spectra, evaluation of single peaks as biomarkers, and forming a panel of biomarkers using multiple peaks. In each of these steps there are a number of issues that are common to other biomarker discovery analyses such as those targeting gene expressions. A major biological criticism against the MALDI-TOF-MS-based tumor biomarker discovery⁴ is low sensitivity of the MALDI-TOF-MS approach relative to the likely concentrations of tumor markers in biological specimens, such as sera. While biostatisticians and bioinformaticians cannot support or refute this criticism directly, empirical evaluations of the criticism can be performed.

In our opinion, a critical step is the preanalysis processing of the data, for which increasing efforts for improving existing methods are seen, for example, by M.D. Anderson Cancer Center's Bioinformatics Group. In view of the point of criticism that protein markers are likely to exist in lower concentrations than MALDI-TOF MS can detect, the preanalysis processing needs to identify peaks with high

sensitivity (a low number of false negative peaks), possibly sacrificing specificity (a moderate number of false positive peaks). Specifically, peak identification approaches that compromise sensitivity in an effort to increase specificity could miss critical but ambiguous peaks lowering the highest discriminatory power of peaks. Since the classification step should be able to separate important peaks from many irrelevant peaks, it would not hurt to compromise specificity to some extent in the preanalysis processing step. Note also that a marker peak specific to a subset of cases may be absent in the rest of cases as well as in controls, which also makes considerations of peak identification strategies complex.

In addition to improving the preanalysis processing step, we see two pressing issues. One is to enable concise and effective comparisons of analysis results across studies. Diamandis⁴ pointed out the inconsistency in the results of several prostate cancer biomarker studies that used SELDI-TOF MS. The inconsistency potentially can be explained fully by the reporting of best peaks in each study. Intensity measurements of peaks in MALDI-TOF mass spectra are highly correlated and contain appreciable noise. If each study reports only its top few peaks, one cannot compare results across studies effectively and completely. A method is needed for concise and effective comparisons of analysis results across studies.

Another important issue is the consideration of problem-specific matters in each tumor marker discovery project (Zhang).²⁷ For example, in prostate cancer biomarker discovery, it is critical to address issues of false labeling of biopsy-missed cancer cases. Validation studies conducted by the NCI's Early Detection Research Network²⁸ are designed with this problem in mind. In breast cancer biomarker discovery, it may be critical to consider heterogeneity across hormone-receptor-defined subtypes. Etiological studies have shown consistent opposite directions of risk-factor associations between cases expressing both estrogen and progesterone receptors, the majority of breast cancer, and those expressing neither receptor, a subtype more commonly seen in younger cases. Biomarker expressions across subtypes may be not only heterogeneous but also opposite in directions relative to controls. This makes a global search of breast cancer biomarkers challenging and necessitates models for interactions with subtypes. Problem-specific innovations by biostatistics and bioinformatics expertise have the potential to provide important contributions to the discovery of proteomic tumor biomarkers.

REFERENCES

1. Tanaka K, Waki H, Ido Y, Akita S, Yoshida Y, Yoshida T. Protein and polymer analyses up to m/z 100,000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*. 2:151–3, 1988.
2. Karas M, Hillenkamp F. Laser desorption of proteins with molecular masses exceeding 10,000 Daltons. *Analytical Chemistry*, 60:2299–301, 1988.
3. Mann M, Talbo G. Developments in matrix-assisted laser desorption/ionization mass spectrometry. *Current Opinion in Biotechnology*, 7:11–9, 1996.
4. Diamandis EP. Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. *Journal of the National Cancer Institute* 96: 353–6, 2004.

5. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y. Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute*, 93:1054–61, 2001.
6. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies: I. Principles. *American Journal of Epidemiology*, 135:1019–28, 1992.
7. Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics*, 20:777–85, 2004.
8. Sorace JM, Zhan M. A data review and reassessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics*, 4:24, 2003.
9. Baggerly KA, Morris JS, Wang J, Gold D, Xiao LC, Coombes KR. A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics*, 3:1667–72, 2003.
10. Yasui Y, Pepe M, Thompson ML, Adam BL, Wright G Jr, Qu Y, Potter JD, Winget M, Thornquist M, Feng Z. A data-analytic strategy for protein-biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics*, 4:449–63, 2003.
11. Yasui Y, McLerran D, Adam BL, Winget M, Thornquist M, Feng Z. An automated peak-identification/calibration procedure for high-dimensional protein measures from mass spectrometers. *Journal of Biomedicine and Biotechnology*, 4:242–8, 2003.
12. Tibshirani R, Hastie T, Narasimhan B, Soltys S, Shi G, Koong A, Le QT. Sample classification from protein mass spectrometry, by “peak probability contrasts.” *Bioinformatics*, 20:3034–44, 2004.
13. Coombes KR, Fritsche HA Jr, Clarke C, Chen JN, Baggerly KA, Morris JS, Xiao LC, Hung MC, Kuerer HM. Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clinical Chemistry*, 49:1615–23, 2003.
14. Randolph T, Yasui Y. Multiscale processing of mass spectrometry data. University of Washington Biostatistics Working Paper Series #1063. Berkeley Electronic Press. 2004.
15. Mallat S. *A Wavelet Tour of Signal Processing*. 2nd ed. New York: Academic Press, 1998.
16. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.
17. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44:837–45, 1988.
18. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.
19. Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, 64:479–98, 2004.
20. Friedman JH, Hastie T, Tibshirani R. Additive Logistic regression: a statistical view of boosting. *Annals of Statistics*, 28, 337–407, 2000.
21. Ruczinski I, Kooperberg C, LeBlanc M. Logic regression. *Journal of Computational and Graphical Statistics*, 12:475–511, 2003.
22. Vapnik VN. *The Nature of Statistical Learning Theory*. New York:Springer, 1995.
23. Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:955–74, 1998.
24. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer, 2001.

25. Djavan B, Mazal P, Zlotta A, Wammack R, Ravery V, Remzi M, Susani M, Borkowski A, Hruby S, Boccon-Gibod L, Schulman CC, Marberger M. Pathological features of prostate cancer detected on initial and repeat prostate biopsy: results of the prospective European Prostate Cancer Detection study. *Prostate*, 47:111–7, 2001.
26. Yasui Y, Pepe M, Hsu L, Adam BL, Feng Z. Partially supervised learning using an EM-boosting algorithm. *Biometrics*, 60:199–206, 2004.
27. Zhang Z. Combining multiple biomarkers in clinical diagnostics: a review of methods and issues. In: Diamandis EP, Fritsche HA, Lilja H, Chan DW, Schwartz MK, eds. *Tumor Markers: Physiology, Pathobiology, Technology, and Clinical Applications*. Washington, DC: AACCC Press, 2002.
28. Srivastava S, Kramer BS. Early detection cancer research network. *Laboratory Investigation*, 80:1147–8, 2000.

25 Statistical Approaches for High Dimensional Data Derived from High Throughput Assays: A Case Study of Protein Expression Levels in Lung Cancer

Yu Shyr

CONTENTS

25.1	Introduction	457
25.2	Class Comparison: Variable Selection	460
25.2.1	Information-Theoretic Score	460
25.2.2	Weighted Gene Analysis	461
25.2.3	Significance Analysis of Microarrays	461
25.2.4	Permutation t-Test	462
25.2.5	Functional Data Analysis (FDA)	462
25.2.6	Inconsistency of Variable Selection	463
25.3	Class Prediction	464
25.3.1	Compound Covariate Method (CCM)	464
25.3.2	Weighted Flexible Compound Covariate Method	465
25.3.3	Leave-One-Out Cross-Validated Class Prediction Model	466
25.3.4	Cluster Analysis: How to Interpret	468
25.4	Final Remarks	468
	Acknowledgments	469
	References	469

25.1 INTRODUCTION

The major challenge in high throughput experiments, such as microarray data, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF

MS) data, or surface-enhanced laser desorption/ionization time-of-flight mass spectral (SELDI-TOF MS) data, is that the data are often high dimensional. When the number of dimensions reaches thousands or more, the computational time for the statistical analyses or pattern recognition algorithms can become unreasonable. This can be a problem, especially when some of the features (markers or variables) are not discriminatory. The irrelevant features may also cause a reduction in the accuracy of some algorithms. For example, experiments with a decision tree classifier have shown that adding a random binary feature to standard datasets can deteriorate the classification performance by 5–10%.¹ Furthermore, in many pattern recognition tasks, the number of features represents the dimension of a search space; the larger the number of features, the greater the dimension of the search space, and the harder the problem. With today's technology, the dimensions of microarray and MALDI-TOF data can reach above 40,000 and 100,000 per experiment. One of the challenges in the traditional statistical approaches is that the number of the subjects is much less than the number of the features or variables in the research dataset.

Although the dimension of the data derived from high throughput assays is extremely large, just as in any biomedical investigations, the core issue of high dimensional data analysis is still based on the primary objective(s) of the experiment. In general, there are three objectives in high throughput experiments: class discovery (nonsupervised method), class comparison (supervised methods), and class prediction (supervised methods). The investigator may be interested in one or more of these three objectives, but the statistical considerations for experiment design and statistical data analysis should correspond to the objective(s) of the study. For example, hierarchical clustering techniques may be useful in class discovery but may not be appropriate in class comparison. The objective for class discovery can be to answer the question, "Are there more than three histological groups, i.e., large cell, adenocarcinoma and squamous cell, in non-small cell lung cancer?" The objective for class comparison can be to answer the question, "Is there a set of features, i.e., genes or proteins, that are differentially expressed between patients responding and nonresponding to a therapy?" The objective for class prediction can be to answer the question, "Is there a set of genes or proteins that can predict the 3-year survival status?" For research with very small sample size, such as cell line study or clinical pilot study, the statistical analysis may focus only on the class comparison because the prediction results may not be very informative. In most clinical studies, the class prediction model is followed by a class comparison or feature selection procedure. In this chapter, we will discuss class comparison and prediction methods with an example from Vanderbilt-Ingram Cancer Center Lung Cancer Specialized Programs of Research Excellence (SPORE).²

Lung cancer is a challenging worldwide clinical problem and the leading cause of death from cancer in the United States for both men and women, with an estimated 171,900 new cases and 157,200 deaths in 2003. Its overall incidence is increasing, and despite complex aggressive approaches to treatment and great strides in understanding its biology and causes, corresponding improvements in outcome are not yet apparent. The behavior of individual non-small cell lung cancer (NSCLC) tumors cannot be understood through the analysis of individual or small numbers of genes, so cDNA microarray analysis has been employed with some success to simultaneously investigate thousands of RNA expression levels and begin to identify patterns

associated with biology. However, mRNA expression is poorly correlated with levels of protein expression, and such analysis cannot detect important posttranslational modifications of proteins, such as proteolytic processing, phosphorylation, or glycosylation, all of which are important processes in determining protein function. Accordingly, comprehensive analysis of protein expression patterns in tissues might improve our ability to understand the molecular complexities of tumor cells.

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) can profile proteins up to 50 kilodaltons in size in tissues (see the chapter by Yasui in this volume for more details on MALDI-TOF technology). This technology not only can directly assess peptides and proteins in sections of tumor tissue, but also can be used for high resolution imaging of individual biomolecules present in tissue sections. The protein profiles obtained can contain thousands of data points, necessitating sophisticated data analysis algorithms.

In an investigation at Vanderbilt–Ingram Cancer Center, patients seen at the Vanderbilt University School of Medicine Hospital between March 1998 and July 2002 for NSCLC and metastases to lung were assessed for the Vanderbilt lung SPORE research. A total of 93 tissue samples (79 lung tumors, 14 normal) were studied. The data were divided into a training data set (42 lung tumors, 8 normal) and a test data set (37 lung tumors, 6 normal).

Protein expression profiles were analyzed using a Voyager Elite 2 MALDI-TOF mass spectrometer (PerSeptive Biosystems, Framingham, USA) equipped with a 337 nm nitrogen laser. Briefly, data acquisition was performed in linear mode geometry under 25 kV of accelerating potential and optimized delayed extraction conditions. Each spectrum was the result of 256 laser shots randomly acquired over the surface of the matrix spot. In this analysis, signals in the mass-to-charge range between 2,000 and 25,000 m/z were considered. Data were internally calibrated with peaks from hemoglobin components (α and β chains). A baseline of each spectrum was corrected using Data Explorer software (Applied Biosystems, Foster City, CA), then binned with an algorithm written by the Bioinformatics group of Vanderbilt–Ingram Cancer Center and used for further statistical analyses.

The MALDI-TOF MS peaks were aligned across samples by use of a genetic algorithm parallel search strategy developed specifically for this task. The peaks were binned together such that the number of peaks in a bin from different samples was maximized while the number of peaks in a bin from the same sample was minimized. This multiobjective fitness function was used to identify optimal bins for the MS peaks. The optimal bins were used to define protein variables for statistical analysis. This procedure is preferable to defining variables at each individual mass unit because it accounts for nonlinear shifts of the peaks between samples. After the data preprocessing procedure, such as baseline correction, normalization, and binning, there were around 1,000 peaks in the final dataset for the statistical analysis.

The primary statistical challenges of analysis of protein expression patterns are to identify a set of MS peaks (or proteins) that are differentially expressed between different classes and to develop predictive models of the statistical relationships between multivariate protein expression data and the clinical features. In this chapter, the mutual-information scoring (Info Score), weighted gene analysis (WGA), significance analysis of microarrays (SAM), permutation t -test or F-test, Kruskal–Wallis

test, Fisher's exact test, and functional data analysis have been reviewed for identifying the MS peaks that are differentially expressed between different classes. One of the challenges for these analyses is that the MS peaks identified by different methods may not be the same. The class prediction model may be used to examine the significances of the MS peaks. Hedenfalk et al.³ successfully applied the t -test based compound covariate method to class prediction analysis for BRCA1+ vs. BRCA1-. We review a recently proposed weighted flexible compound covariate method (WFCCM)^{4,5} based on Info Score, WGA, SAM, Kruskal-Wallis test, permutation test, and Fisher's exact test. We conclude with suggestions for general ideas of analyzing MALDI-TOF MS protein expression data.

25.2 CLASS COMPARISON: VARIABLE SELECTION

Most classification and pattern recognition methodologies are limited by the number of independent variables that can be assessed simultaneously. Thus, a useful subset of independent variables must be selected. This is referred to as the feature or variable subset selection problem. The mutual-information scoring (Info Score), weighted gene analysis (WGA), significance analysis of microarrays (SAM), and permutation t -tests or F-tests have been successfully applied in several RNA and proteomic expression experiments for determining the genes/proteins that are differentially expressed between classes; see, for example, Hedenfalk et al.,³ Tusher et al.,⁶ Beer et al.,⁷ Yamagata et al.,⁸ and Yanagisawa et al.² These methods provide the significance, the score or the rank of the genes/proteins of the multi-dimensional data, which forms the basis for subset selection.

25.2.1 INFORMATION-THEORETIC SCORE

The Info Score is an information-theoretic score, which was introduced by Ben-Dor, Friedman, and Yakhini.⁹ It uses a rank-based scoring system and combinatorial permutation of sample labels to produce a rigorous statistical benchmarking of the overabundance of features whose differential expression pattern correlates with sample type, such as tumor (+) versus normal (-). Let N denote the number of tissues, consisting of p tissues from class P and q tissues from Class Q with expression levels g_{jp} and g_{jq} , respectively. P and Q could denote tumor and normal tissue samples, for example. We define g_j as an N -vector of expression level and rank vector v_j of g_j as an N -vector of $\{+, -\}$:

$$v_j = \begin{cases} + & \text{if } g_j \in P \\ - & \text{if } g_j \in Q \end{cases} \quad (25.1)$$

For example, if the expression levels for feature or gene/protein j are $\{1, 2, 3, 5, 6, 7, 11, 14\}$ in class P and $\{4, 8, 9, 10, 12, 13, 15\}$ in class Q , then $v_j = \{+, +, +, -, +, +, -, -, -, +, -, -, +, -\}$. Note that the rank vector v_j captures the essence of the differential expression profile of g_j . If g_j is underexpressed in class P , then the positive entries of v_j are concentrated in the left-hand side of the vector, and the negative entries are concentrated at the right-hand side.

The Info score of a rank vector $\mathbf{v}_j$ is defined as

$$Info(\mathbf{v}_j) = \min_{x,y=\mathbf{v}} \left\{ \left(\frac{|x|}{|\mathbf{v}|} \right) Ent(x) + \left(\frac{|y|}{|\mathbf{v}|} \right) Ent(y) \right\} \quad (25.2)$$

where $|\bullet|$ = cardinality and $Ent(\bullet)$ is the entropy of $\bullet$ defined by

$$Ent(\bullet) = H(\phi) = -\phi \log_2(\phi) - (1 - \phi) \log_2(1 - \phi) \quad (25.3)$$

and ϕ denotes the fraction of positive entries in $\bullet$. Entropy can be viewed as measurement of the degree of disorder or uncertainty in a system. Using the same example above, the best partition with respect to the Info Score for feature or gene/protein j is

$$Info(\mathbf{v}_j) = \frac{7}{15} H\left(\frac{6}{7}\right) + \frac{8}{15} H\left(\frac{2}{8}\right) = 0.71 \quad (25.4)$$

The range of the Info Score is between 0 and 1; the smaller Info Scores indicate stronger evidence of the different expression profiles of two classes.

In summary, the Info Score uses a rank-based scoring system and combinatorial permutation of sample labels to produce a rigorous statistical benchmarking of the overabundance of genes/proteins whose differential expression pattern correlates with sample type. Genes or proteins may be ranked on the basis of Info Score.

25.2.2 WEIGHTED GENE ANALYSIS

Weighted gene analysis (WGA) was proposed by Bittner and Chen¹⁰ and can be defined as follows. For a given two-class setting, P versus Q , a discriminative weight (score) for each feature j can be evaluated by

$$w_j = \frac{d_B}{(f_1 d_{wP} + f_2 d_{wQ} + \alpha)} \quad (25.5)$$

where d_B is the center-to-center Euclidean distance between the two classes, d_{wg} is the average Euclidean distance among all sample pairs within class g , $g = P, Q, L$ and M are the number of sample pairs in classes P and Q , respectively, $f_1 = L/(L + M)$ and $f_2 = M/(L + M)$, and α is a small constant, such as 0.01, to prevent zeros in the denominator.

The range of the w_j is between 0 and ∞ , and the bigger w_j scores indicate stronger evidence of the different expression profiles of two classes.

25.2.3 SIGNIFICANCE ANALYSIS OF MICROARRAYS

Significance analysis of microarrays (SAM)⁶ is a statistical technique for finding significant genes in a set of microarray experiments. It is a method for identifying genes on a microarray with statistically significant changes in expression, developed in the context of an actual biological experiment. SAM assigns a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements to estimate the percentage of genes identified by chance, the false positive

rate (FPR) (see the chapters by Matsui and Hsu et al. in this volume). It is based on the analysis of random fluctuations in the data. In general, the signal-to-noise ratio is decreased with decreasing gene expression. The relative difference $d(j)$ in gene expression is

$$d(j) = \frac{\bar{x}_P(j) - \bar{x}_Q(j)}{s(j) + s_0} \quad (25.6)$$

where $\bar{x}_P(j)$ and $\bar{x}_Q(j)$ are defined as the average levels of expression for gene j in class P and Q , respectively, and

$$s(j) = \sqrt{a(\Sigma_p[x_p(j) - \bar{x}_P(j)]^2 + \Sigma_q[x_q(j) - \bar{x}_Q(j)]^2)} \quad (25.7)$$

where $x_p(j)$ and $x_q(j)$ are defined as the expression level for gene j in classes P and Q , respectively, and Σ_p and Σ_q are summations of the expression measurements in classes P and Q , respectively, $a = (1/p + 1/q)(p + q - 2)$, and p and q are the numbers of measurements in classes P and Q , respectively.

The distribution of $d(j)$ should be independent of the level of gene expression. At low expression levels, the variance in $d(j)$ can be high because of small values of $s(j)$. To ensure that the variance of $d(j)$ is independent of gene expression, SAM adds a small positive constant s_0 to the denominator.

The coefficient of variation (CV) of $d(j)$ can be computed as a function of $s(j)$ in moving windows across the data. The value of s_0 can be chosen to minimize the coefficient of variation. The bigger $|d(j)|$ scores indicate the stronger evidence of the different expression profiles of two classes. Although SAM is designed for microarray data analysis, it can be applied to the protein profile data analysis because the data structures of these two technologies are similar. The data transformation method may be different between microarray and protein profile data. In general, the log 2 transformation is recommended for microarray data analysis, while the log 10 transformation of the data is commonly applied to proteomic expression data.

25.2.4 PERMUTATION t -TEST

The permutation t -test¹¹ is a strategy for establishing differences in gene/protein expression pattern between classes. In the permutation t -test, the standard t statistic is computed on the log-expression ratios of each feature in order to analyze the variation between classes. Then, labels of the classes are randomly permuted among the specimens and the t statistic for each feature j in the permuted data set is computed. This process is repeated 10,000 or more times³. Finally, a critical value of the t statistic, such as 0.999, is determined for each feature based on the empirical distribution of t from permuted data sets for the feature j . If the t statistic for a feature j in the original labeling of specimens is larger than its critical value, the feature is deemed differentially expressed between the two groups and is considered as significantly different.

25.2.5 FUNCTIONAL DATA ANALYSIS (FDA)

Protein spectra from MALDI-TOF are an instance of functional data. That is, the intrinsic structure of the data can be viewed as a continuous function of mass per

charge. The motivation for FDA methods is to treat the entire measured function as a single observation rather than as a closely spaced sequence of measurements. The advantage of this approach is that we may incorporate into the analysis methods continuity and smoothness constraints believed present in the data generating process. Recently developed statistical methods for functional data are summarized in Ramsay and Silverman.¹² These methods include, among others, functional data extensions to linear models, principal components analysis, canonical correlation, and discriminant analysis. The goals of FDA are similar to those in other areas of statistics: to gain insight into sources of pattern and variation regarding the data generating mechanism. Like the other methods, the features can be ranked by a test, such as the F-statistic, which indicates the significance of the features.

25.2.6 INCONSISTENCY OF VARIABLE SELECTION

We have reviewed seven methods for the variable or feature selection process in class comparison. There are several other existing methods that can do a similar job, such as the REML-based mixed model,¹³ or the threshold number of misclassification score (TNoM).⁹ There are also more new methods coming out for class comparison based on high dimensional data. More new methods will be developed as microarray, SELDI-TOF, or MALDI-TOF research advances.

Each of these methods can generate a list of features based on their significances or scores. The question is, “Do the results from these methods agree with each other?” The answer is “No.” Based on the nature of the development of each method, the results from each method will most likely produce different rankings of features. Figure 25.1 shows the results from the Vanderbilt lung cancer SPORC microarray study (Yamagata et al., 2003) based on four methods — permutation *t*-test, SAM, WGA, and Info Score. In this study, total RNAs were extracted from 26 tumors (24 lung tumors, 1 breast, and 1 sarcoma) and 3 normal tissues. The 5,088 cDNAs were prepared from sequence-verified colons of Research Genetics to represent 4,749 unique genes including 2,126 expressed sequence tags (ESTs) and 2,587 non-ESTs. The investigators would like to

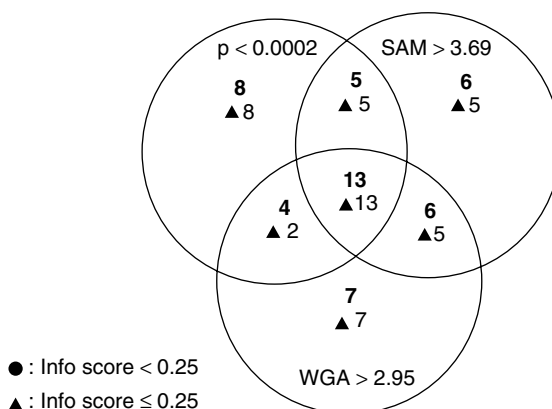


FIGURE 25.1 Variable selection results of Vanderbilt Lung Cancer SPORC Microarray Study.

generate a set of genes that performed differently between non-small cell lung cancer tumors and normal tissues. We picked the top 30 genes that performed most differently between these two classes using each of these four statistical methods. There were 49 “winner” genes based on the union of all four methods. Among these winner genes, 21 (42.9%), 15 (30.6%) and 13 (26.5%) genes were selected as winners by two, three, and all four methods, respectively. The similar inconsistency of variable selection was found in the Vanderbilt lung cancer SPORE MALDI-TOF study.² The results of the selection were influenced by sample size, variation of the data, percent of missing data, as well as the outliers/extreme values of the data within each class. In practice, it is not a bad idea to use more than one method to generate the feature list. The investigators may focus on the features selected by all methods first.

25.3 CLASS PREDICTION

Class prediction methods can be used to examine the goodness of the set of features identified in the class comparison step. There are two types of class prediction methods: those based on the training data set and those based on the test data set. For high throughput data analysis, it is highly recommended that class prediction models be applied to the test data set because they may easily overfit the training data set. In addition, the sample size of the blinded/test data set probably should be comparable to the sample size of the training data set if the sample size of the training data set is very small. On the other hand, if the sample size of the training data set is large, such as several hundred samples, a training:test set ratio of $k:1$, where $k > 1$, may be appropriate.

25.3.1 COMPOUND COVARIATE METHOD (CCM)

Hedenfalk et al.³ successfully applied the compound covariate method (CCM)¹⁴ to class prediction analysis for BRCA1+ versus BRCA1– breast cancer tumor types. This predictor is built in two steps. First, a standard two-sample t -test is performed to identify genes with significant differences (at level α , Hedenfalk et al.³ picked $\alpha = 0.0001$) in log-expression ratios between the two tissue classes. Second, the log-expression ratios of differentially expressed genes are combined into a single compound covariate for each tissue sample; the compound covariate is used as the basis for class prediction. The compound covariate for tissue sample i is defined as

$$c_j = \sum_j t_j x_{ij} \quad (25.8)$$

where t_j is the t -statistic for the two-group comparison of classes with respect to gene j , x_{ij} is the log-ratio measured in tissue sample i for gene j , and the sum is over all differentially expressed genes.

The CCM reduces the data dimension from $N \times J$ to $N \times 1$ to where N is the total number of samples and the J is total number of study features. We can view CCM as the overall score of each tissue sample, which combines information of all important features from one statistical method.

25.3.2 WEIGHTED FLEXIBLE COMPOUND COVARIATE METHOD

The weighted flexible compound covariate method (WFCCM)^{2,4,5,8} is an extension of the compound covariate method, which allows consideration of more than one statistical analysis method in the compound covariate. Before we apply the WFCCM, it is important to make sure that the sign of each statistical method is consistent. For example, the sign of the t -statistic and SAM are always consistent, but WGA scores are always positive because its scoring system is based upon Euclidean distance. Therefore, multiplying a^{-1} to the WGA scores for all the features that have negative scores in SAM or t -statistic is the first step for applying the WFCCM. The second step is to select the winners with all statistical methods. We may arbitrarily pick genes from each statistical method, such as the top 1% or top 100 features, or we may use some significant information to select features from the statistical methods, such as p -value < 0.0001 for t -statistic, p -value < 0.01 for REML-based mixed effect models, or SAM > 3.5 .

The WFCCM for tissue sample i is defined as

$$WFCCM(i) = \sum_j \left[\sum_k (ST_{jk} W_k) \right] [W_j] x_{ij} \quad (25.9)$$

where x_{ij} is the log-ratio measured in tissue sample i for feature j . ST_{jk} is the standardized statistic/score of feature j , such as the standardized SAM score, for statistical analysis method k . W_k is the weight of method k , which can be determined as

$$W_k = (1 - CCM \text{ misclassification rate}_k) \quad (25.10)$$

where CCM misclassification rate _{k} stands for the misclassification rate of the compound covariate method for statistical analysis method k for $k = 1, \dots, K$. W_j is the weight of feature j , which can be determined as

$$W_j = \sum_k \frac{V_{jk}}{K} \quad (25.11)$$

where $V_{jk} = 1$ if the feature j is selected as the winner in method k and, 0 otherwise. If feature j is selected by all methods then $W_j = 1$.

The W_k and W_j can also be determined by other methods. For example, we may assign $W_k = 1$ for all K methods used in the variable selection stage if we believe they perform equally well. We may also modify W_j as

$$W_j = \left[\left(\sum_k \frac{V_{jk}}{K} \right) (1 - Info \text{ Score}_j) \right] \quad (25.12)$$

In this case, if feature j is selected by all methods and the $Info \text{ Score}_j = 0$, then $W_j = 1$.

The WFCCM also reduces the data dimension from $N \times J$ to $N \times 1$. We can certainly view WFCCM as the overall score of each tissue sample, which combines all information of all important features from several statistical methods.

25.3.3 LEAVE-ONE-OUT CROSS-VALIDATED CLASS PREDICTION MODEL

Cross-validation is a method for estimating generalization error based on resampling. Many classification methodologies do not have straightforward procedures for testing null versus alternate hypotheses using p -values or likelihood ratio tests. The primary measure of success for these methods is the error or misclassification rate.¹⁵ Because many classification methodologies are susceptible to overfitting the data, it is common to base success on the predictive ability of models. This is accomplished through cross-validation. For example, in a procedure known as 10-fold cross-validation, the data are divided into 10 equal parts and the model is developed using 9/10 of the data and tested for its ability to predict the remaining 1/10 of the data. This is repeated a total of 10 times with 10 training sets and 10 test sets. A model that has a low misclassification rate (<5% per group) for prediction in 80% of the test sets may be considered statistically significant. Statistical significance of a classification model can be judged using misclassification rates and cross-validation. When the sample size is small, the misclassification rate can be assessed using the leave-one-out cross-validated (LOOCV) class prediction method. LOOCV is one specific type of cross-validation.

The LOOCV is processed in four steps in the Vanderbilt lung cancer SPORE study training set. First, the WFCCM is applied to calculate the single compound covariate for each tissue sample based on the significant MS peaks. Second, one tissue sample is selected and removed from the dataset, and the distance between the two tissue classes for the remaining tissue samples is calculated. Third, the removed tissue sample is classified based on the closeness of the distance of the two tissue classes using, for example, the k -nearest neighbor approach, with $k = 2$, or the midpoint of the means of the WFCCM for the two classes as the threshold. Fourth, steps 2 and 3 are repeated for each tissue sample. To determine whether the accuracy for predicting membership of tissue samples into the given classes as measured by the number of correct classifications is better than the accuracy that may be attained for predicting membership into random grouping of the tissue samples, we created 5,000 random data sets by permuting class labels among the tissue samples. Cross-validated class prediction is performed on the resulting data sets, and the percentage of permutations that result in as few or fewer misclassifications than the original labeling of samples can be reported. If less than 5% of permutations result in as few or fewer misclassifications, the accuracy of the prediction of the given classes is considered significant. Therefore, this rate may be considered as the p -value for the class prediction model. We recently have succeeded in applying the WFCCM class prediction analysis method to the preliminary data generated by the Vanderbilt lung cancer SPORE study.

In the training data set, the perfect WFCCM class-prediction models based on 82, 91, 23, and 20 differentially expressed MS peaks were found to classify lung tumor versus normal lung tissue samples, primary NSCLC lung cancer versus normal lung tissue samples, primary NSCLC lung cancer versus other type of lung tumor tissue samples, and adenocarcinoma versus squamous cell carcinoma tissue samples, respectively. Table 25.1 shows the results from the Vanderbilt lung cancer study. WGA, SAM, Kruskal–Wallis test, Fisher’s exact test, and permutation t -test were included in the compound covariate using cut-off points 2.0, 3.5, $p < 0.0001$, $p < 0.0001$, and $p < 0.0001$, respectively. We selected $W_k = 1$ for all methods, and W_j using (25.12).

TABLE 25.1
WFCCM Class Prediction Model in Training Data Set

Classification (Sample Size)	No. of Diff. Expressed Peaks	No. of Misclassified Samples (%)	Prob. of Random Permutations with Misclassifications
All samples			
Normal lung (8) vs. tumor (42)	82	0 (0)	<0.0001
Normal lung (8) vs. primary NSCLC (34)	91	0 (0)	<0.001
Primary NSCLC (34) vs. other type of lung tumor (7)	23	0 (0)	<0.001
Non-small cell lung cancer			
Adeno (14) vs. squamous (15)	20	0 (0)	<0.0001
Adeno (14) vs. large (5)	20	0 (0)	<0.01
Squamous (15) vs. large (5)	12	0 (0)	<0.01
Nodal involvement negative (20) vs. positive (14)	2	5 (14.7)	<0.05

TABLE 25.2
WFCCM Class Prediction Model in Test Data Set

Classification (Sample Size)	No. of Diff. Expressed Peaks	No. of Misclassified Samples (%)	Percentage of Correctly Classified Samples (95% C.I.)
All samples			
Normal lung (6) vs. Lung tumor (37)	82	0	100 (92, 100)
Normal lung (6) vs. primary NSCLC (32)	91	0	100 (91, 100)
Primary NSCLC (32) vs. other type of lung tumor (5)	23	0	100 (91, 100)
Non-small cell lung cancer			
Adeno (13) vs. squamous (16)	20	0	100 (88, 100)
Adeno (13) vs. large (3)	20	1	93.8 (70, 99)
Squamous (16) vs. large (3)	12	0	100 (82, 100)
Nodal metastasis negative (25) vs. positive (7)	2	8	75.0 (57, 88)

We also applied the WFCCM to a set of blinded test samples described in Section 25.1. Table 25.2 shows the results of the analyses. In general, the model performed reasonably well with the average correct prediction rate in the blinded test data set exceeding 93% except for the prediction of negative versus positive nodal metastasis. The latter case is not surprising because there were only two MS peaks reaching the selection criteria. The lower bounds of the 95% confidence interval of the mean percentage of the correctly classified samples are all above 50%; therefore, the results of correct prediction rate in the blinded testing data sets are not likely to happen by chance.

25.3.4 CLUSTER ANALYSIS: HOW TO INTERPRET

Cluster analysis is the tool for grouping objects of similar kind into respective categories; it is not designed for the class prediction problem. Cluster analysis is widely used in high dimensional data analysis. However, the results from cluster analysis for high dimensional data must be interpreted very carefully, especially when supervised statistical tools are applied in the feature selection procedure. A common mistake is to use the cluster analysis for an objective of the class prediction based on the pre-selected winner features.

The hierarchical agglomerative method is one of the commonly used cluster methods that can be very useful to display graphically (with a dendrogram) the relations contained in a resemblance matrix. However, clustering algorithms may present problems when dealing with larger sample size, such as 45,000 genes or 100,000 m/z peaks, because trees become uninterpretable. In addition, at the highest levels of the hierarchy, the branches are totally dependent on the lower links, so they are less reliable as partitions. In general, almost all hierarchical agglomerative algorithms are safe when the interest is placed on the small, low levels of the hierarchy, and they become unsafe at high levels. Hierarchical algorithms will normally differ in the solutions given at those high levels.

Figure 25.2 shows the results from the agglomerative hierarchical clustering algorithm for clustering lung tumors and normal lung tissues. The number of features in the cluster analysis was selected as described in Section 25.3.2. The average linkage algorithm was applied to calculate the distance between the clusters. All the lung tumor tissues clustered together, as did the normal lung tissues. The results looked very promising but might only be used to reconfirm the MS peaks that performed differently between the two classes. Having a perfect or near perfect cluster result was expected if we applied cluster analysis after we selected the features that performed differently using any of the supervised methods. It is important to note that we could not apply these results to any class discovery conclusion.

25.4 FINAL REMARKS

In summary, the statistical class comparison and class prediction analyses for high throughput data may focus on the following steps:

1. Select important feature patterns that perform differently among the study groups.
2. Use the class prediction model based on the WFCCM or other methods to verify if the features selected in step 1 have statistically significant prediction power on the training samples.
3. Apply the prediction model generated from step 2 to a set of test samples for validation.
4. Employ the agglomerative hierarchical clustering algorithm to investigate the pattern among the significant discriminator features as well as the biologic status.

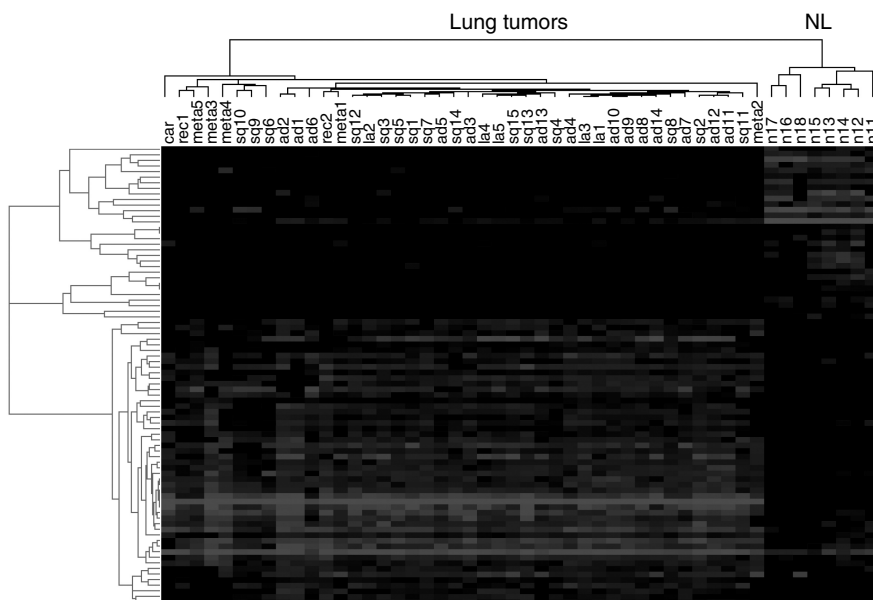


FIGURE 25.2 (See color insert following page 458.) Results from the agglomerative hierarchical clustering algorithm.

Although perfect or near perfect predictors are reported in our study, some statistical limitations need be addressed. First, because the study sample size is small, a larger scale study to confirm our findings is necessary. In addition, the number of features reported in this paper is not based on the smallest number of features that could discriminate the classes but is based on the statistical evidence. The possibility of achieving similar misclassification rates based on different subsets of peaks certainly exists. In addition, the near perfect discrimination obtained using the agglomerative hierarchical clustering is not surprising as it uses covariates that were themselves chosen to have maximal discriminating power.

ACKNOWLEDGMENTS

This work was supported in part by the Lung Cancer SPORE (Special Program of Research Excellence) (P50 CA90949) and Cancer Center Support Grant (CCSG) (P30 CA68485). The authors thank Dr. David Carbone, principal investigator of Vanderbilt Lung Cancer SPORE, for permission to use the study data for the illustration.

REFERENCES

1. Witten I.H. and Frank E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, San Francisco, 1999.

2. Yanagisawa K., Shyr Y., Xu B.J., Massion P.P., Larsen P.H., Whit B.C., Roberts J.R., Gonzalez A., Nadaf S., Moore J.H., Caprioli R.M., and Carbone, D.P., Tumor proteomic patterns predict classification and tumor behavior in human non-small cell lung cancer, *Lancet*, 362:433–439,2003.
3. Hedenfalk I., Duggan D., Chen Y., Radmacher M., Bittner M., Simon R., Meltzer P., Gusterson B., Esteller M., Kallioniemi O.P., Wilfond B., Borg A., and Trent J., Gene-expression profiles in hereditary breast cancer, *N Engl J Med*, 344:539–548,2001.
4. Shyr Y., Analysis and interpretation of array data in human lung cancer using statistical class-prediction model, in *Program/Proceeding Supplement of AACR meeting April, 2002*, San Francisco, 2002,41–42.
5. Shyr Y. and Kim K.M., Weighted flexible compound covariate method for classifying microarray data, in *A Practical Approach to Microarray Data Analysis*, Berrar, D.P., Dubitzky W., and Granzow M., eds., Kluwer Academic Publishers, Norwell, 2003, 186–200.
6. Tusher V.G., Tibshirani R., and Chu G., Significance analysis of microarrays applied to the ionizing radiation response, *Proc Natl Acad Sci USA*, 98:5116–5121,2001.
7. Beer D.G., Kardia S.L., Huang C.C., Giordano T.J., Levin A.M., Misek D.E., Lin L., Chen G., Gharib T.G., Thomas D.G., Lizyess M.L., Quick R., Hayasaka S., Taylor J.M., Iannettoni M.D., Orringer M.B., and Hanash S., Gene-expression profiles predict survival of patients with lung adenocarcinoma, *Nat Med*, 8:816–824,2002.
8. Yamagata N., Shyr Y., Yanagisawa K., Edgerton M., Dang T.P., Gonzalez, A., Nadad S., Larsen P., Roberts J.R., Nesbitt J.C., Jensen R., Levy S., Moore J.H., Minna J.D., and Carbone D.P., A training-test approach to the molecular classification of resected non-small cell lung cancer, *Clin Cancer Res*, 9:4695–4704,2003.
9. Ben-Dor A., Friedman N., and Yakhini Z. Scoring genes for relevance. Tech Report AGL-2000-13, Agilent Labs, Agilent Technologies, 2000.
10. Bittner M. and Chen Y., Statistical methods: Identification of differentially expressed genes by weighted gene analysis, 2001. Available at <http://www.nejm.org/general/content/supplemental/hedenfalk/index.html>.
11. Radmacher M.D., Simon R. Statistical methods: generating gene lists with permutation F and *t* tests, 2001. Available at <http://www.nejm.org/general/content/supplemental/hedenfalk/index.html>.
12. Ramsay J.O. and Silverman B.W., *Functional Data Analysis*, Springer-Verlag, New York, 1997.
13. Wolfinger R.D., Gibson G., and Wolfinger E.D., Assessing gene significance from cDNA microarray expression data via mixed models, *J Comput Biol*, 8:625–637,2001.
14. Tukey J.W., Tightening the clinical trial, *Control Clin Trials*, 14:266–285,1993.
15. Ripley B., *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, 1996.

26 Spatial Modeling of Multilocus Data

David V. Conti, Daniel O. Stram, John Molitor, Paul Marjoram, and Duncan C. Thomas

CONTENTS

26.1	Introduction	471
26.2	Genotype-Based and Haplotype-Based Penetrance Models	472
26.3	Likelihoods and Case-Control Ascertainment	474
26.4	Multiple Comparisons, Sparse Data, and Bayesian Smoothing or Clustering	475
26.5	Relationship to Coalescent Methods	477
26.6	Haplotype Sharing Methods	478
26.7	Other Issues	479
	References	482

26.1 INTRODUCTION

As we gain further knowledge regarding the underlying etiology of cancers and as technological advances continue for characterizing genetic variation, the study of the genetic component to cancer susceptibility and progression is becoming more salient. There has been some success at narrowing regions harboring cancer-influencing loci with linkage analysis. For example, Witte et al.¹ present compelling evidence for the identification of several loci for prostate cancer risk and tumor aggressiveness. However, there is an increasing need to use genetic association studies to evaluate genes conferring a modest risk or genes that interact with many of the already identified environmental risk factors for cancer. Genetic association studies broadly seek to demonstrate a relation between genetic variation and disease. Subsumed in these approaches are aims to determine regions of the genome that may influence genes, i.e., gene hunting, and to estimate relative risk and location of specific causal polymorphisms, i.e., gene characterization.

Because of their high density throughout the human genome — about one every 200 base pairs,² their relative ease of high-throughput genotyping,³ and the rapidly

growing catalogs of variants,⁴ single nucleotide polymorphisms (SNPs) are ideally suited for linkage disequilibrium (LD) mapping of unknown genes and testing candidate gene associations. Conventionally in a genotype-level analysis, each SNP is investigated independently for association. As an alternative, one may also examine the association of several SNPs jointly by focusing on haplotypes or the arrangement on chromosomes of the alleles at each SNP. Because most molecular techniques characterize only genotypes, haplotypes are unobserved, and haplotype-level analysis requires the estimation of haplotype frequencies for the population or the estimation of the most likely pair of haplotypes for each individual. However, the recent recognition that LD tends to be concentrated in blocks of limited haplotype diversity has sparked great interest in the use of haplotypes for both gene hunting and gene characterization. For example, in a multiethnic sample, Haiman et al.⁵ used a haplotype-based approach within *CYP19* to characterize linkage disequilibrium (LD) and haplotype patterns, and to investigate the involvement of several haplotypes with breast cancer. We review here a broad range of methods that have been proposed for the analysis of multilocus data in association studies, with particular emphasis on haplotype and Bayesian methods. We begin, however, with a general likelihood-based framework for inference on genotype and haplotype associations in case-control studies using unrelated individuals.

26.2 GENOTYPE-BASED AND HAPLOTYPE-BASED PENETRANCE MODELS

Let Y_i denote the phenotypes of a sample of $i = 1, \dots, n$ subjects. For most purposes, we will restrict attention to case-control data, where $Y = 1$ indicates a case and $Y = 0$ a control for cancer status such as breast, prostate, or colon cancer. However, the generalized linear model we propose can be applied to most any phenotype by appropriate selection of a link function, thus allowing for investigation of measures of aggressiveness, number of polyps, etc. Suppose we also observe for each subject a vector of $m = 1, \dots, M$ SNP genotypes $\mathbf{G}_i = (G_{im})$ where $G_{im} = 0, 1, 2$ indicates the number of copies of a particular variant allele, conventionally, the rarer allele. Using standard logistic regression approaches, we might consider a penetrance model of the form

$$\text{logit } \Pr(Y_i = 1 \mid \mathbf{G}_i) = \beta_0 + \sum_m \beta_m Z(G_{im}) + \dots \quad (26.1)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_M)$ is a vector of corresponding log odds ratios for the M SNPs. $Z(G)$ denotes some coding of the genotypes incorporating any assumptions about the genetic mode of transmission. For example, to model a dominant genetic mode we may code both the homozygote genotype for the variant alleles and the heterozygote genotype as $Z(G) = 1$ and the homozygous wild-type genotype as $Z(G) = 0$. In Equation (26.1) the ellipsis (...) indicates the possibility of including covariates or additional locus-by-locus or gene-environment interaction terms. We call this a genotype-based model and note that such an approach does not require knowledge

of the genotypes' phase, i.e., how the alleles are arranged on chromosomes. We contrast this approach with a haplotype-based model of the form

$$\text{logit Pr}(Y_i = 1 \mid \mathbf{H}_i) = \gamma_0 + \sum_{h \sim \mathcal{H}} Z(\mathbf{H}_i) \gamma_h + \dots \quad (26.2)$$

where $\mathbf{H}_i = (h_{i1}, h_{i2})$ designates a pair of haplotypes h for individual i in the space $\mathcal{H}$ of all haplotypes represented in the population. $Z(\mathbf{H}_i)$ denotes some coding for the pair of haplotypes for each individual with corresponding vector of haplotype log odd ratios, $\gamma = (\gamma_h)_{h \in \mathcal{H}}$. Often $Z(\mathbf{H}_i)$ simply represents variables indicating the count of the most likely haplotypes for individual i . For the haplotype analysis with k_m alleles at each of M markers, there are $\prod_{m=1}^M k_m$ possible haplotypes in the space $\mathcal{H}$. For example, when looking at M SNPs where $k_m = 2$ for all SNPs, there are 2^M potential haplotypes. Since these haplotypes usually cannot be determined with certainty from the observed genotypes, the full prospective likelihood is given by

$$L_{(P)}(\boldsymbol{\gamma}, \mathbf{q}) = \Pr(\mathbf{Y}, \mathbf{G}) = \prod_{i=1}^n \sum_{\mathbf{h} \sim \mathbf{G}_i} \Pr(Y_i \mid \mathbf{H}_i = \mathbf{h}; \boldsymbol{\gamma}) \Pr(\mathbf{H}_i = \mathbf{h} \mid \mathbf{q}) \quad (26.3)$$

where the summation is over the set $\mathbf{h} \sim \mathbf{G}_i$ of haplotype pairs that are compatible with each individual's observed genotypes. In addition to the haplotype relative risks $\boldsymbol{\gamma}$, this likelihood is also a function of the population haplotype frequencies $\mathbf{q} = (q_h)_{h \in \mathcal{H}}$, assuming the population is in Hardy–Weinberg (H–W) equilibrium, $\Pr(\mathbf{H} \mid \mathbf{q}) = q_{h_1} q_{h_2}$.

For a reasonable number of loci or haplotypes, the model can be fitted using the E–M algorithm.^{6–9} This is commonly done using a two-stage procedure in which $\mathbf{q}$ is first estimated from the controls (or in the total sample ignoring the phenotypes) and then treated as fixed in a score test of $H_0: \boldsymbol{\gamma} = \mathbf{0}$ based on Equation (26.3). The justification of combining cases and controls in this test is that under H_0 , there is no difference in haplotype frequencies between the two groups, but this is of course violated if the aim is estimation of $\boldsymbol{\gamma}$ rather than hypothesis testing. Stram et al.¹⁰ have extended this basic E–M approach to a single-stage joint estimation of $\boldsymbol{\gamma}$ and $\mathbf{q}$. Their E-step entails calculation of the expectation of the number of copies $N_h(\mathbf{G}_i)$ of each haplotype $h \in \mathcal{H}$ given both $\mathbf{G}_i$ and Y_i , conditional on the current estimates of the parameters, and the M-step entails maximization of the complete-data likelihood for $\boldsymbol{\gamma}$ and $\mathbf{q}$ using these expectations.

This calculation can become unwieldy if there are a large number of possible haplotypes. One possible solution is simply to limit the analysis to the subset of those haplotypes with some arbitrary minimal estimated population frequency, such as greater than 5%. No such restriction is needed, however, if Markov chain Monte Carlo (MCMC) methods are used.^{11–13} Here, the basic idea is to alternate between sampling from $[\mathbf{H}_i \mid \mathbf{G}_i, Y_i; \boldsymbol{\gamma}, \mathbf{q}]$, $[\boldsymbol{\gamma} \mid Y_i, \mathbf{H}_i]$, and $[\mathbf{q} \mid \mathbf{H}_i]$, where $[\bullet \mid \bullet]$ denotes the respective full conditional distributions. The first of these updates can easily be done by a Metropolis–Hastings step, proposing a switch of the alleles at a single locus or a contiguous segment. The remaining updates use standard MCMC moves. We have implemented similar MCMC approaches to haplotype assignment in pedigrees, where a change to any individual's haplotype must be propagated through the rest of the pedigree and the Hastings ratio is now also a function of the recombination

fractions. Yet another approach uses estimating equations methods.¹⁴ In theory, the number of haplotypes can be quite large as the number of SNPs increases. However, in practice the number that are actually observed in human populations over reasonably small regions tends to be modest.^{15,16} We will revisit this issue when we consider haplotype block structure below.

26.3 LIKELIHOODS AND CASE-CONTROL ASCERTAINMENT

For logistic models with fully observed covariates, such as the genotype model in Equation (26.1), there is a well known equivalence between the prospective likelihood $\Pr(\mathbf{Y} | \mathbf{G}, n(\mathbf{Y}))$ (where $n(\mathbf{Y})$ denotes the number of subjects with $Y = 1$) and the retrospective likelihood $\Pr(\mathbf{G} | \mathbf{Y})$.^{17,18} Proof of this relationship is complex and relies on two key observations: first, that the odds ratio for $[\mathbf{Y} | \mathbf{G}]$ is the same as the odds ratio for $[\mathbf{G} | \mathbf{Y}]$, and second, that consistent estimates of the parameters of this odds ratio can be obtained by allowing the distribution of $\mathbf{G}$ to be unspecified or to depend on parameters that are independent of the odds ratio. For further discussion in the context of family studies see.^{19,20} This is the fundamental justification for the appropriateness of the prospective likelihood for case-control studies, even though it is Y that is sampled and G that is observed.

For models with incompletely observed covariates, such as the haplotype model (Equation (26.2)), this equivalence no longer holds. The likelihood given in Equation (26.3) is appropriate for cohort studies but not for case-control studies without additional correction for the differential sampling fractions of cases and controls. Stram et al.¹⁰ have shown that naive application of the cohort likelihood to case-control data can lead to substantial bias in the population haplotype frequency estimates, essentially because the high risk haplotypes are overrepresented in a case-control sample. Because the estimates of $\boldsymbol{\gamma}$ and $\mathbf{q}$ are not independent, this can also lead to some bias in $\hat{\boldsymbol{\gamma}}$, although the magnitude of this bias is generally small and depends upon how accurately the haplotypes can be predicted from the genotypes.

This bias can be eliminated in three ways. The standard conditional likelihood, $\Pr(\mathbf{Y} | \mathbf{G}, n(\mathbf{Y}))$, can be computationally daunting for large strata because of the need to sum over all possible permutations of Y with the same $n(\mathbf{Y})$ as that observed.²¹ A simpler approach, but still prospective, is to condition on the marginal probability of each subject being ascertained,

$$L_{(PA)}(\boldsymbol{\gamma}, \mathbf{q}) = \Pr(\mathbf{Y} | \mathbf{G}, Asc)$$

$$= \prod_{i=1}^n \frac{\sum_{h \sim G_i} \pi_{Y_i} \Pr(Y_i | \mathbf{H}_i = \mathbf{h}; \boldsymbol{\gamma}) \Pr(\mathbf{H}_i = \mathbf{h} | \mathbf{q})}{\sum_{h \sim G_i} [\pi_1 \Pr(Y_i = 1 | \mathbf{H}_i = \mathbf{h}; \boldsymbol{\gamma}) + \pi_0 \Pr(Y_i = 0 | \mathbf{H}_i = \mathbf{h}; \boldsymbol{\gamma})] \Pr(\mathbf{H}_i = \mathbf{h} | \mathbf{q})}$$

where π_1 and π_0 are the case and control sampling fractions respectively. This likelihood requires knowledge of the sampling fractions π_1 and π_0 , which are readily available for a nested case-control study within a cohort but would require knowledge of the

marginal disease rate in the population for a conventional population-based case-control study. Alternatively, the retrospective likelihood

$$L_{(R)}(\boldsymbol{\gamma}, \mathbf{q}) = \Pr(\mathbf{G}_i | Y_i) = \frac{\sum_{\mathbf{h} \sim \mathbf{G}_i} \Pr(Y_i | \mathbf{H}_i = \mathbf{h}; \boldsymbol{\gamma}) \Pr(\mathbf{H}_i = \mathbf{h} | \mathbf{q})}{\sum_{\mathbf{h} \in \mathbf{H}} \Pr(Y_i | \mathbf{H}_i = \mathbf{h}; \boldsymbol{\gamma}) \Pr(\mathbf{H}_i = \mathbf{h} | \mathbf{q})}$$

can be used. Epstein and Satten²² show that this likelihood can be expressed in a simpler manner if parameterized in terms of the haplotype frequencies in controls and the odds ratios; this also avoids having to assume the haplotypes are in H-W equilibrium. Kraft and Thomas²³ compare the performance of the retrospective and (ascertainment-corrected) prospective likelihood approaches in the case of fully observed covariates and conclude that the latter often provides a more efficient estimator of $\boldsymbol{\beta}$, despite the need to address the additional $\mathbf{q}$ parameters. Whether this is true in the case of incompletely observed covariates remains to be studied.

26.4 MULTIPLE COMPARISONS, SPARSE DATA, AND BAYESIAN SMOOTHING OR CLUSTERING

A number of authors have considered the relative efficiency of genotype-based and haplotype-based models.^{8,24–28} Because LD between individual SNPs is typically too high to fit multiple logistic models involving all SNPs simultaneously, most of these authors have compared a haplotype-based model (typically a multiple degree of freedom test) against the best-fitting single-SNP models with Bonferroni correction for multiple comparisons. Of course, it is also necessary to sum over the unknown phases in any haplotype-based test. Results of such comparisons have been variable, depending upon the specific tests compared, assumptions regarding the extent of LD, and the causal model for penetrance. If LD is high and a single haplotype contains the causal variant, a haplotype-based test can be more powerful, despite the larger degrees of freedom and the uncertainty of phase. On the other hand, if LD is low (so that phase is difficult to infer) or if there is a single causal variant among the set of available SNPs and its effect is spread out over a number of different haplotypes, then a genotype-based test can be more powerful.

Additionally, if there are many genotypes and/or haplotypes, we are confronted with the related problems of multiple comparisons and sparse data, for which Bayesian shrinkage estimators offer a natural solution. The idea is to allow the estimates of effect (the $\boldsymbol{\beta}$ s or $\boldsymbol{\gamma}$ s) to borrow information from similar genotypes or haplotypes. This problem is very similar to that considered in the Bayesian spatial smoothing and spatial clustering literature, with diverse applications such as spatial modeling of disease rates in environmental epidemiology.^{29–32}

To incorporate the dependencies among SNPs, Conti and Witte³³ have proposed a hierarchical genotype-based model in which each $\boldsymbol{\beta}_m$ is estimated univariately in the first-level model. A second-level regression model of the form $\boldsymbol{\beta} \sim N(\mathbf{X}' \boldsymbol{\alpha}, \boldsymbol{\Sigma}(x))$ is used to smooth the $\hat{\boldsymbol{\beta}}$ s where $\mathbf{X}$ is a matrix of prior covariates for each SNP and $\boldsymbol{\Sigma}(x)$ is a covariance matrix that includes spatial dependencies among SNPs or could depend upon the location x of the postulated causal variant. A recent extension of this

approach presents a genotype-level analysis to jointly model the SNPs and includes a pairwise phase term aimed at capturing the underlying haplotype structure.³⁴ Thus, rather than selecting between genotype- or haplotype-level approaches, the method frames the analysis of multilocus data in a model selection paradigm, the aim being to determine which SNPs, phase terms, and linear combinations best describe the relation between genetic variation and disease. In related work, Bayesian model selection or model averaging techniques^{35,36} have been used to allow for our uncertainty about whether particular SNPs or interaction terms between candidate genes within a biologic pathway should be included in the model.³⁷

In a haplotype-based model, we wish to exploit the notion that structurally similar haplotypes in the neighborhood of a disease predisposing locus are more likely to harbor the same susceptibility allele and hence to have similar γ s. Thomas et al.³⁸ and Molitor et al.³⁹ considered a conditional autoregressive (CAR) model of the form

$$\gamma \sim N(\mathbf{0}, \sigma^2 \mathbf{I} + \tau^2 \mathbf{W})$$

where $\mathbf{W}$ is a matrix of similarities of each pair (h, k) of haplotypes, such as the length $L_{hk}(x)$ of the segment shared identical by state (IBS) surrounding a candidate mutation location x . Standard MCMC methods are used to update γ , σ^2 , and τ^2 . The location x is updated by a Metropolis–Hastings move, proposing either a small random walk in the neighborhood of the current location or an entirely new location anywhere in the region. One may assume a uniform prior for x across the region of interest or place a higher prior density in regions suspected of harboring a candidate mutation.

More recently, we have been considering the Potts⁴⁰ and Voronoi⁴¹ spatial clustering models of the form

$$\text{logit Pr}(Y_i = 1 \mid \mathbf{H}_i) = \delta_0 + \delta_{c_{h1}} + \delta_{c_{h2}}$$

where c_h denotes the cluster to which haplotype h belongs and δ_c is a relative risk parameter common to all haplotypes h assigned to cluster c . As a further generalization, one could allow the risk for a given haplotype h to be a function of the risk for which the particular haplotype is assigned δ_{c_h} , i.e., $\gamma_h \sim N(\delta_{c_h}, \sigma^2)$. For the Potts model, each element c_h of the assignment vector $\mathbf{c}$ indicates the cluster label to which haplotype h belongs $\{c_h = 1, 2, \dots, C\}$. The probability of a realization of the entire assignment vector $\mathbf{c}$ is

$$\text{Pr}(\mathbf{c}) = \frac{\exp[\psi \sum_{hk} W_{hk}(x) I(c_h = c_k)]}{\Psi[C, \psi, \mathbf{W}(x)]}$$

where $\Psi[C, \psi, \mathbf{W}(x)]$ is a normalizing constant equal to the sum of the numerator over all possible partitions. In this formula, the term inside the exponential is the weighted sum over all like-labeled haplotypes where the weights are determined according to the appropriate haplotype similarity metric. The interaction parameter ψ is nonnegative with $\psi = 0$ corresponding to independent allocations uniformly on

the labels $\{1, 2, \dots, C\}$. The degree of spatial smoothing increases with ψ . This leads to a simple expression for the full conditional distribution of the cluster assignment for any particular haplotype,

$$\Pr(c_h = c) = \frac{\exp[\psi \sum_{hk} W_{hk}(x) I(c_h = c)]}{\sum_{c'=1}^C \exp[\psi \sum_{hk} W_{hk}(x) I(c_k = c')]} \quad (1)$$

but updates of the parameters C , ψ , and x involve the normalizing constant Ψ , which can be quite complex. In the Voronoi model, haplotypes are assigned deterministically to the cluster containing the nearest ancestral haplotype A_c ,⁴² thereby avoiding the need to consider any normalizing constants. As in the CAR model, MCMC methods are used to update c , ψ , x , and A_c . Reversible jump MCMC methods⁴³ are needed to update the number of clusters C , but these calculations are greatly simplified by integrating out the $\mathcal{Y}$ as described by Denison and Holmes.³² The current version of the algorithm deals only with phase-known haplotype data. While versions for dealing with unphased data are in development, the algorithm has been successfully applied to an analysis of data from *Arabidopsis thaliana* which, as a selfing organism, is for all practical purposes phase-known.⁴⁴

One can combine the ideas of Potts and Voronoi approaches by using the Voronoi model but assigning haplotypes to centers probabilistically via a Potts modeling approach. Specifically, one can express the probability of haplotype assignment to a cluster c given the cluster center A_c as

$$\Pr(c_h = c | A_c) = \frac{\exp[\psi \sum_h W_h(x, A_c)]}{\sum_{c'=1}^C \exp[\psi \sum_h W_h(x, A_{c'})]} \quad (2)$$

Here the normalizing constant is simply a sum over the number of clusters, not the sum over all possible haplotype allocations. W_h is the weight that corresponds to haplotype h belonging to cluster c with cluster ancestor haplotype A_c and cluster mutation location x_c , as defined above. By incorporating an extra step of estimating latent cluster centers, we can reduce the dimensionality of the parameter space.

26.5 RELATIONSHIP TO COALESCENT METHODS

We view the spatial smoothing and spatial clustering approaches as a relatively simple empirical approximation to the more formal coalescent methods that have received a great deal of attention as a possible approach to LD mapping.^{45–49} The coalescent, introduced by Kingman,⁵⁰ describes the probability distribution for the tree structure describing the ancestral relationships between a present-day sample of haplotypes and their most recent common ancestor (MRCA), together with the associated times for coalescence of each pair of branches and the mutation rate parameter. Kuhner et al.⁵¹ and Griffiths and Tavaré^{52,53} describe MCMC methods for inference in coalescent models. Whereas the coalescent assumes that all the variation between present-day haplotypes is due to mutation, a generalization known as

the ancestral recombination graph (ARG) allows for both mutation and recombination, leading to a graph topology in which joins (moving backwards in time) represent coalescence events and splits represent recombinations.^{54–56} Effective MCMC samplers for the ARG have remained elusive, although there has been some progress.^{56–59} We are currently exploring forms of rejection sampling known as approximate Bayesian computation^{60,61} that avoids the need to compute the likelihood for any given realization of the topology by comparing the closeness of random data sampled under that realization to the observed data. For further discussion of coalescent methods, see Nordborg⁶² and Stephens.⁶³

In a similar vein to our Bayesian spatial clustering model, several authors^{64–67} have proposed maximum likelihood methods based on the idea of ancestral haplotype reconstruction from a sample of present-day case and control haplotypes. In these papers, the ancestral haplotype(s) were treated as parameters to be estimated along with the various population parameters (mutation and recombination rates, mutation locations, penetrances, etc.). Morris et al.^{68,69} instead used MCMC methods to allow the ancestral haplotypes to be treated as latent variables to be sampled over rather than maximized out.

26.6 HAPLOTYPE SHARING METHODS

An important prediction of coalescent models is that pairs of cases would tend to be more closely related than pairs of controls, while case-control pairs would be even more distantly related on average.⁷⁰ This observation underlies a class of LD mapping methods known as haplotype sharing, in which one searches for locations where apparently-unrelated case pairs tend to have more sharing of haplotypes than other pairs. Although this approach has been used informally for meiotic mapping for years,⁷¹ the approach was first formalized by Te Meerman and Van Der Meulen⁷² as a permutation test of a haplotype sharing statistic (HSS), compared with a null distribution obtained by randomly permuting the cases and controls. In general, a broad class of HSS can be formulated as $H = \sum_{hk} L_{hk}(x) D_{hk}$ where $L_{hk}(x)$ is the length of the segment surrounding location x that is shared *IBS* by haplotypes h and k , and $D_{hk} = (Y_h - \mu)(Y_k - \mu)$ is a score for the phenotypic similarity of the individuals carrying these haplotypes, the sum being taken over all pairs of haplotypes from apparently unrelated individuals. A number of variants of this general approach have been discussed, including the haplotype sharing correlation,⁷³ the maximum identity length contrast (MILC),^{74–76} and various other tests.^{77–83}

In their simplest forms, these various approximations of the coalescent treat each of the present-day haplotypes as independently descended from their respective ancestral haplotype(s); this is known as a star-shaped genealogy (or equivalently, the absence of cryptic relatedness.^{84,85} McPeck and Strahs⁶⁴ and Morris et al.⁶⁹ also discuss extensions of their approach to allow for dependency within the sets of descendants from each ancestral haplotype. Our spatial clustering models also have assumed a star-shaped genealogy. One way around this difficulty would be to propagate the hierarchical clustering, in which $\gamma_h \sim N(\delta_c, \sigma^2)$ and $\delta_c \sim N(\eta_{d_c}, \tau^2)$ (where d_c are clusters of clusters), and so on. Reversible jump MCMC methods could be used to allow the number of levels and branches of the hierarchy to be unknown.

A more elegant approach, however, incorporates the estimated time T_{hk} to a common ancestor for each pair of haplotypes within a cluster based on their shared length L_{hk} thereby exploiting the notion that haplotypes that are more similar to each other are more likely to be closely related. We therefore allow each haplotype to have its own γ_h with covariance $\text{cov}(\gamma_h, \gamma_k | T_{hk}) = \sigma_c^2 \exp(-\phi T_{hk})$. We define the priors $T_{hk} \sim \Gamma(1, 1/(2N))$ and $L_{hk}(x) \sim \Gamma(2, 2T_{hk})$ where N is the effective population size. Thus, the marginal covariance can be shown to be

$$\text{cov}(\gamma_h, \gamma_k | c_h = c_k) = \left(\frac{4\phi\sigma_c^2}{N} \right) \left(\frac{L_{hk}}{\phi + 1/(2N) + 2L_{hk}} \right)$$

Such methods may provide an alternative approach to the problem of cryptic stratification in which a population consists of a number of subpopulations that are not readily distinguishable by self-reported race/ethnicity but in fact differ both in terms of allele frequencies at the candidate locus and baseline rates of disease (see the reviews by Thomas and Witte,⁸⁶ Wacholder et al.,⁸⁷ and Cardon and Palmer.⁸⁸ A variety of methods involving the use of a panel of unlinked markers to infer the latent population structure have been proposed.^{89–92} We believe such approaches can be readily incorporated into our Bayesian spatial clustering models.

Our implementation of the CAR and Voronoi models currently assumes phase-known haplotype data are available, as in a set of transmitted (case) and nontransmitted (control) haplotypes derived from case-parent triads. Extension to the phase-unknown case appears straightforward, however, simply involving the kinds of Metropolis–Hastings moves discussed above. The acceptance probabilities would now be a function of the relative likelihood of the cluster assignments of the new and old haplotypes rather than the marginal population haplotype frequencies. Even more appealing is the possibility of exploiting coalescent methods, in which haplotypes might be assigned conditional on the sampled individuals' Y and G together with the current topology of the coalescent tree or ARG, and then the topology and associated parameters are updated conditional on the current assignment of haplotypes.

26.7 OTHER ISSUES

In genetic association studies, there are still several unresolved questions as to the optimal approach of relating genetic variation to disease via the evaluation of multilocus data. Association studies have traditionally been classified into direct or indirect approaches. For the direct approach, putative disease susceptibility polymorphisms with *a priori* evidence for a functional role are tested individually for association. In contrast, the indirect approach to association studies of either unrelated individuals or case-parent trios relies on linkage disequilibrium (LD) across a set of polymorphisms to capture the underlying genetic mechanism for disease. This is true of candidate gene approaches or, as is becoming increasingly possible, whole genome association scans.

However, the boundary between these two approaches is recently clouding. Because epidemiological and biological evidence for a functional role of any given polymorphism varies, many researchers are opting to supplement known function

variants with a set of SNPs designed both to increase our chances of finding and evaluating a true causal variant and to capture the underlying genetic variation. In this approach difficulties originate in how to best represent the underlying genetic variation in the sample of interest. Here we have focused on the statistical issues for the analysis of multilocus data given a set of SNPs. Additional issues surround selection of the best method to choose the set of SNPs — especially in the situation of an ethnically diverse sample in which varying evolutionary histories lead to diverse patterns of LD.

In this context, the recent discovery that the pattern of LD throughout the genome appears to be concentrated in relatively short blocks makes haplotype-based selection of SNPs much more attractive.^{15,16,93} Haplotype blocks are regions with a limited number of distinct haplotypes separated by regions of low LD. The latter could reflect either hot spots of recombination or ancient recombination events.^{94–97} Most importantly, it implies that to identify all the polymorphisms within a region that are relevant to disease one could simply select a subset of haplotype-tagging SNPs (htSNPs) that in combination are sufficient to predict all the common haplotypes within each block. Thus, in using a subset of SNPs via the haplotype structure as a proxy for an underlying disease mechanism, current research programs of either a particular set of genes,^{98–100} or the entire human genome, such as the HapMap Project,^{101–103} rely on several assumptions. First, they assume that a large proportion of the susceptibility of a given disease is due to a small number of common variants, each contributing a modest effect. Second, these common variants are either SNPs or are thought to arise at rates no greater than the rate that new SNPs appear. Third, the rates at which SNPs arise and become common are slow enough so that most of the carriers of a common variant today come from a single ancestor. Finally, the local recombination rates are low enough that carriers of the disease variant will also tend to carry a similar set of SNPs, reflecting the shared common ancestry.

These projects aim to characterize the haplotype block structure in an effort to reduce the number of SNPs that would be required to capture the essential genetic variation, as suggested by Risch and Merikangas.¹⁰⁴ These projects are continually evolving, and there is some ambiguity as to what information will actually be available, e.g., allele frequencies or haplotype structure by ethnic group, and whether or not that information is applicable to most study samples. Additionally, there is some controversy surrounding the common-disease common-variant hypothesis^{105,106} and the number of SNPs actually required. Very recently, Carlson et al.¹⁰⁷ compared the results of their own SNP discovery efforts from resequencing to SNPs already deposited in the dbSNP database at <http://www.ncbi.nlm.nih.gov/SNP>. By extrapolation from their results for 50 genes, they estimated that all 2.7 million known SNPs in that database are barely sufficient in whites and quite insufficient in African Americans to predict well the majority of their newly discovered SNPs. Unfortunately, the statistic they used to determine whether a given SNP discovered by their sequencing was well predicted by the dbSNP database was the maximum of simple pairwise R^2 s, i.e., for each new SNP discovered in the sequencing, the maximal pairwise R^2 between that SNP and each dbSNP was computed. Simple pairwise R^2 may be a very poor measure of the actual predictability either of common SNPs that may be of even more importance or of common haplotypes based on any given set of htSNPs.

A number of other approaches to htSNP selection have been described,^{15,16,108–114} but the number of htSNPs that are needed in whole-genome association studies is presently not known. Our approach¹¹⁵ for determining the number and identity of htSNPs is based on maximizing the minimum across all common haplotypes of the statistic R^2_h , the expected correlation between the true number $N_h(G_i)$ of copies of haplotype h for each individual and its predicted value given the full subset of SNPs. This approach fully exploits the multivariate pattern of LD but requires reasonably good estimation of all the haplotype frequencies in a particular block and is dependent on H–W equilibrium. An initial simulation¹¹⁵ found good behavior of the R^2_h statistic when the true state of nature was that of quite limited haplotype diversity within a block. In this simulation, however, 70 individuals of a particular ethnic group were used to characterize the haplotypes, which is considerably more than the 23 African Americans and 24 whites sequenced by Carlson et al.¹⁰⁷ Further investigation performing quasi-empirical power calculations using the HapMap ENCODE region of chromosome 2 found that SNP marker densities of one common SNP per 5kb appears to provide htSNPs that give large case-control studies adequate power to reject the null hypothesis.¹¹⁶ A computer program to compute R^2_h and find minimal sets of htSNPs using this criteria is available from our website at <http://www-rcf.usc.edu/~stram/tagSNPs.html>.

Chapman et al.¹⁰⁸ describe an alternative approach based on choosing the subset of SNPs that maximizes the minimum across loci m of R^2_m , the multiple correlation between G_{im} and the subset, on the assumption that the causal variant being sought is likely to be in the dataset; this measure again is distinguished from the pairwise R^2 used by Carlson et al.¹⁰⁷ in that it exploits the multivariate pattern of LD. If a set of haplotype frequency estimates are available, e.g., by running an E-M algorithm, a formal calculation of the expected squared correlation R^2_m can be performed¹¹⁵ rather than estimating this squared correlation directly from a regression analysis of non-htSNP genotypes upon htSNP genotypes. This formal calculation is dependent upon the assumption of H–W equilibrium for the haplotypes. It is possible that the formal calculation is a more efficient estimate of R^2_m if H–W equilibrium does indeed hold.

In general, studies utilizing haplotype structure to reduce the number of SNPs via a set of tagging SNPs have relied on a two-stage design approach.^{5,117} Due to prohibitive costs of genotyping a large number of SNPs in a large case-control sample, these studies first use a subset of the data to select a set of tagging SNPs from a larger set of all known polymorphism within the candidate region of interest. Then, only these tagging SNPs are genotyped in the larger main study sample. While there has been little research into the efficiency of these two-stage studies, we have recently been investigating optimal designs in terms of both capturing the genetic variation via a subset of SNPs and the corresponding cost efficiency.¹¹⁸ While results indicate that the two-stage design can be very cost efficient compared with simply genotyping all known SNPs on all study subjects, care must be taken in choosing the subsample for tag SNP selection. In addition, the optimal design further depends on the goals of the analysis (e.g., estimation or hypothesis testing, estimating relative risks, or localizing a causal variant), the analysis paradigm used (haplotype- or genotype-based models), and the representativeness of the subsample in capturing the genetic variation that may potentially exist in the larger ethnically diverse main study.

Issues surrounding the ability of a subset of SNPs to adequately represent the gene-disease relation and the corresponding two-stage design approaches may become obsolete as genotyping costs plummet and it becomes increasingly cost effective to genotype all known polymorphisms for all study subjects. As this future becomes reality, there will be increasing reliance on statistical models with the ability to evaluate numerous genetic polymorphisms while incorporating knowledge of the underlying genetic structure and addressing problems of multiple comparisons and sparse data. The computational challenges are formidable and an area deserving further attention. We are continuing to pursue these issues from a Bayesian perspective relying heavily on the ideas of spatial analysis, which we believe offers the prospect of providing a unified framework that adequately allows for the uncertainty in our characterization of the relation of genetic variation to cancer susceptibility and progression.

REFERENCES

1. Witte JS, Suarez BK, Thiel B, et al. Genome-wide scan of brothers: replication and fine mapping of prostate cancer susceptibility and aggressiveness loci. *Prostate* 2003;57:298–308.
2. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nat Genet* 2003;33 Suppl:228–37.
3. Oliphant A, Barker DL, Stuelplnagel JR, Chee MS. BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques* 2002;Suppl:56–8, 60–1.
4. Sachidanandam R, Weissman D, Schmidt SC, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001;409:928–33.
5. Haiman CA, Stram DO, Pike MC, et al. A comprehensive haplotype analysis of CYP19 and breast cancer risk: the multiethnic cohort. *Hum Mol Genet* 2003;12:2679–92.
6. Chiano M, Clayton D. Fine genetic mapping using haplotype analysis and the missing data problem. *Ann Hum Genet* 1998;62:55–60.
7. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 1995;12:921–7.
8. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 2002;70:425–34.
9. Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered* 2002;53:79–91.
10. Stram DO, Leigh Pearce C, Bretsky P, et al. Modeling and e-m estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Hum Hered* 2003;55:179–90.
11. Niu T, Qin ZS, Xu X, Liu JS. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet* 2002;70:157–69.
12. Liu JS, Sabatti C, Teng J, Keats BJB, Risch N. Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res* 2001;11:1716–24.
13. Stephens M, Donnelly P. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 2003;73:1162–9.
14. Zhao LP, Li SS, Khalid N. A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *Am J Hum Genet* 2003;72:1231–50.

15. Daly M, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. *Nat Genet* 2001;29:229–32.
16. Patil N, Berno AJ, A HD, et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001;294:1719–23.
17. Anderson JA. Separate sample logistic discrimination. *Biometrika* 1972;59:19–35.
18. Prentice R, Breslow N. Retrospective studies and failure time models. *Biometrika* 1978;65:153–8.
19. Whittemore AS, Halpern J. Logistic regression of family data from retrospective study designs. *Genet Epidemiol* 2003;25(3):177–89.
20. Whittemore A. Logistic regression of family data from case-control studies. *Biometrika* 1995;82:57–67.
21. Greenland S, Morgenstern H, Thomas DC. Considerations in determining matching criteria and stratum sizes for case-control studies. *Int J Epidemiol* 1981;10:389–92.
22. Epstein MP, Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet* 2003;73:1316–29.
23. Kraft P, Thomas DC. Bias and efficiency in family-matched gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. *Am J Hum Genet* 2000;66:1119–31.
24. Akey J, Jin L, Xiong M. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet* 2001;9:291–300.
25. Bader JS. The relative power of SNPs and haplotype as genetic markers for association tests [comment]. *Pharmacogenomics* 2001;2:11–24.
26. Long AD, Langley CH. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res* 1999;9:720–31.
27. Morris RW, Kaplan NL. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol* 2002;23:221–33.
28. Morris AP, Whittaker JC, Balding DJ. Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. *Am J Hum Genet* 2004;74:945–53.
29. Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 1987;43:671–81.
30. Knorr-Held L, Rasser G. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* 2000;56:13–21.
31. Mollie A, Richardson S. Empirical Bayes estimates of cancer mortality rates using spatial models. *Stat Med* 1991;10:95–112.
32. Denison DG, Holmes CC. Bayesian partitioning for estimating disease risk. *Biometrics* 2001;57:143–9.
33. Conti DV, Witte JS. Hierarchical modeling of linkage disequilibrium: genetic structure and spatial relations. *Am J Hum Genet* 2003;72:351–63.
34. Conti DV, Gauderman WJ. SNPs, Haplotypes, and model selection in a candidate gene region: the SIMPlE analysis for multilocus data. *Genetic Epidemiology* 2004;27:429–41.
35. Madigan D, Raftery AE. Model selection and accounting for model uncertainty in graphical models using Occam's window. *J Am Statist Assoc* 1994;89:1335–46.
36. Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. *J Am Statist Assoc* 1997;92:179–91.
37. Conti DV, Cortessis V, Molitor J, Thomas DC. Bayesian modeling of complex metabolic pathways. *Hum Hered* 2003;56(1–3):83–93.
38. Thomas DC. Bayesian models and Markov chain Monte Carlo methods. *Genet Epidemiol* 2001;21 (Suppl 1):S660–S1.
39. Molitor J, Marjoram P, Thomas D. Application of Bayesian spatial statistical methods to analysis of haplotypes effects and gene mapping. *Genet Epidemiol* 2003;25:95–105.

40. Green P, Richardson S. Hidden Markov models and disease mapping. *J Am Stat Assoc* 2002;97:1055–70.
41. Voronoi. Nouvelles applications des parametres continus a la theorie des formes quadratiques. *J Reine Angew Math* 1908;34:198–287.
42. Molitor J, Marjoram P, Thomas D. Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *Am J Hum Genet* 2003;73:1368–84.
43. Green P. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *J Roy Statist Soc, ser B* 1995;82:711–32.
44. Hagenblad J, Tang C, Molitor J, et al. Haplotype structure and phenotypic associations in the chromosomal regions surrounding two *Arabidopsis thaliana* flowering time loci. *Genetics* 2004;168:1627–38.
45. Hill WG, Weir BS. Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am J Hum Genet* 1994;54:705–14.
46. Kaplan N, Hill W, Weir B. Likelihood methods for locating disease genes in non-equilibrium populations. *Am J Hum Genet* 1995;56:18–32.
47. Xiong M, Guo S. Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. *Am J Hum Genet* 1997;60:1513–31.
48. Zollner S, von-Haeseler A. A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. *Am J Hum Genet* 2000;66:615–28.
49. Graham J, Thompson EA. Disequilibrium likelihoods for fine-scale mapping of a rare allele. *Am J Hum Genet* 1998;63:1517–30.
50. Kingman JFC. On the genealogy of large populations. *J Appl Prob* 1982;19A:27–43.
51. Kuhner M, Yamato J, Felsenstein J. Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. *Genetics* 1995;140:1421–30.
52. Griffiths R, Tavaré S. Simulating probability distributions in the coalescent. *Theor Pop Biol* 1994;46:131–59.
53. Griffiths R, Tavaré S. Computational methods for the coalescent. In: Donnelly P, Tavaré S, eds. *Population Genetics and Human Evolution*. Berlin: Springer Verlag; 1997:165–82.
54. Hudson R. Properties of a neutral allele model with intragenic recombination. *Theoret Popn Biol* 1983;23:183–201.
55. Griffiths R, Marjoram P. An ancestral recombination graph. In: Donnelly P, Tavaré S, eds. *Progress in Population Genetics and Human Evolution/IMA Volumes in Mathematics and its Applications*. Berlin: Springer Verlag; 1997:257–70.
56. Griffiths R, Marjoram P. Ancestral inference from samples of DNA sequences with recombination. *J Comp Biol* 1996;3:479–502.
57. Kuhner M, Felsenstein J. Sampling among haplotype resolutions in a coalescent-based genealogy sampler. *Genet Epidemiol* 2000;19:S15–S21.
58. Fearnhead P, Donnelly P. Estimating recombination rates from population genetic data. *Genetics* 2001;159:1299–318.
59. Larribe F, Lessard S, Schork NJ. Gene mapping via the ancestral recombination graph. *Theor Popul Biol* 2002;62:215–29.
60. Marjoram P, Molitor J, Plagnol V, Tavaré S. Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci USA* 2003;100:15324–8.
61. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. *Genetics* 2002;162:2025–35.
62. Nordborg M. Coalescent theory. In: Balding DJ, Cannings, C., Bishop, M., eds. *Handbook on Statistical Genetics*. John Wiley & Sons; 2001:179–212.

63. Stephens M. Inference under the coalescent. In: Balding DJ, Bishop M, Cannings C, eds. *Handbook of Statistical Genetics*. John Wiley & Sons; 2001:213–38.
64. McPeck MS, Strahs A. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am J Hum Genet* 1999;65:858–75.
65. Service SK, Lang DW, Freimer NB, Sandkuijl LA. Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *Am J Hum Genet* 1999;64:1728–38.
66. Seltman H, Roeder K, Devlin B. Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. *Am J Hum Genet* 2001;68:1250–63.
67. Seltman H, Roeder K, Devlin B. Evolutionary-based association analysis using haplotype data. *Genet Epidemiol* 2003;25:48–58.
68. Morris A, Whittaker J, Balding D. Bayesian fine-scale mapping of disease loci, by hidden Markov models. *Am J Hum Genet* 2000;67:155–69.
69. Morris AP, Whittaker JC, Balding DJ. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am J Hum Genet* 2002;70:686–707.
70. Markovtsova L, Marjoram P, Tavare S. The age of a unique event polymorphism. *Genetics* 2000;156:401–9.
71. Houwen R, Baharloo S, Blankenship K, et al. Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nat Genet* 1994;8:380–6.
72. Te Meerman G, Van Der Muelen M. Genomic sharing surrounding alleles identical by descent effects of genetic drift and population growth. *Genet Epidemiol* 1997;14:1125–30.
73. Qian D, Thomas D. Genome scan of complex traits by haplotype sharing correlation. *Genet Epidemiol* 2001;21(Suppl 1):S582–S7.
74. Bourgain C, Genin E, Quesneville H, Clerget-Darpoux F. Search for multifactorial disease susceptibility genes in founder populations. *Ann Hum Genet* 2000;64 (Pt 3):255–65.
75. Bourgain C, Genin E, Holopainen P, et al. Use of closely related affected individuals for the genetic study of complex diseases in founder populations. *Am J Hum Genet* 2001;68:154–9.
76. Bourgain C, Genin E, Margaritte-Jeannin P, Clerget-Darpoux F. Maximum identity length contrast: a powerful method for susceptibility gene detection in isolated populations. *Genet Epidemiol* 2001;21(Suppl 1):S560–4.
77. Boon M, Nolte IM, Bruinenberg M, et al. Mapping of a susceptibility gene for multiple sclerosis to the 51 kb interval between G511525 and D6S1666 using a new method of haplotype sharing analysis. *Neurogenetics* 2001;3:221–30.
78. Devlin B, Roeder K, Wasserman L. Genomic control for association studies: a semi-parametric test to detect excess-haplotype sharing. *Biostatistics* 2000;1:369–87.
79. Fischer C, Beckmann L, Marjoram P, te Meerman G, Chang-Claude J. Haplotype sharing analysis with SNPs in candidate genes: the 12 example. *Genet Epidemiol Genetic Analysis Workshop* 2003;24:68–73.
80. Beckmann L, Fischer C, Deck KG, Nolte IM, te Meerman G, Chang-Claude J. Exploring haplotype sharing methods in general and isolated populations to detect gene(s) of a complex genetic trait. *Genetic Epidemiology* 2001;21 Suppl 1:S554–559.

81. Schork N, Thiel B, St. Jean P. Linkage analysis, kinship, and the short-term evolution of chromosomes. *J Exp Zool* 1999;282:133–49.
82. Zhang S, Zhao H. Linkage disequilibrium mapping in population of variable size using the decay of haplotype sharing and a stepwise-mutation model. *Genet Epidemiol* 2000;19:S99–S105.
83. Tzeng JY, Devlin B, Wasserman L, Roeder K. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am J Hum Genet* 2003;72:891–902.
84. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theor Pop Biol* 2001;60:155–60.
85. Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. *Theor Pop Biol* 2001;60:226–37.
86. Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate gene associations? *Cancer Epidemiol Biomarkers Prev* 2002;11:505–12.
87. Wacholder S, Rothman N, Caporaso N. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiologic studies of common polymorphisms and cancer. *Cancer Epidemiol Biomarkers Prev* 2002;11:513–20.
88. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet* 2003;361:598–604.
89. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet* 2000;67:170–81.
90. Satten GA, Flanders WD, Yang Q. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* 2001;68:466–77.
91. Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999;55:997–1004.
92. Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* 2001;20:4–16.
93. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. *Science* 2002;296:2225–9.
94. Wang N, Akey JM, Zhang K, Chakraborty R, Jin L. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am J Hum Genet* 2002; 71:1227–34.
95. Arnheim N, Calabrese P, Nordborg M. Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. *Am J Hum Genet* 2003;73:5–16.
96. Wall JD, Pritchard JK. Assessing the performance of the haplotype block model of linkage disequilibrium. *Am J Hum Genet* 2003;73:502–15.
97. Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* 2003;4:587–97.
98. Bonnen PE, Wang PJ, Kimmel M, Chakraborty R, Nelson DL. Haplotype and linkage disequilibrium architecture for human cancer-associated genes. *Genome Res* 2002;12:1846–53.

99. Crawford DC, Carlson CS, Rieder MJ, et al. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am J Hum Genet* 2004;74:610–22.
100. Reich DE, Cargill M, Bolk S, et al. Linkage disequilibrium in the human genome. *Nature* 2001;411:199–204.
101. Gibbs RA, Belmont JW, Hardenbol P, et al. The International HapMap Project. *Nature* 2003;426:789–96.
102. Couzin J. Genomics: new mapping project splits the community. *Science* 2002;296:1391–3.
103. Cardon LR, Abecasis GR. Using haplotype blocks to map human complex trait loci. *Trends Genet* 2003;19:135–40.
104. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996;273:1616–7.
105. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 2001;69:124–37.
106. Weiss KM, Clark AG. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 2002;18:19–24.
107. Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat Genet* 2003;33:518–21.
108. Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered* 2003;56:18–31.
109. Zhang K, Deng M, Chen T, Waterman MS, Sun F. A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci USA* 2002;99:7335–9.
110. Johnson GC, Esposito L, Barratt BJ, et al. Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001;29:233–7.
111. Sebastiani P, Lazarus R, Weiss ST, Kunkel LM, Kohane IS, Ramoni MF. Minimal haplotype tagging. *Proc Natl Acad Sci USA* 2003;100:9900–5.
112. Thompson D, Stram D, Goldgar D, Witte JS. Haplotype tagging single nucleotide polymorphisms and association studies. *Hum Hered* 2003;56:48–55.
113. Wiuf C, Laidlaw Z, Stumpf MP. Some notes on the combinatorial properties of haplotype tagging. *Math Biosci* 2003;185:205–16.
114. Zhang W, Collins A, Morton NE. Does haplotype diversity predict power for association mapping of disease susceptibility? *Hum Genet* 2004.
115. Stram DO, Haiman CA, Hirschhorn JN, et al. Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum Hered* 2003;55:27–36.
116. Stram DO. Tag SNP selection for association studies. *Genetic Epidemiology* 2004; 27:365–74.
117. Loukola A, Chadha M, Penn SG, et al. Comprehensive evaluation of the association between prostate cancer and genotypes/haplotypes in CYP17A1, CYP3A4, and SRD5A2. *Eur J Hum Genet* 2004;12:321–32.
118. Thomas D, Xie R, Gebregziabher M. Two-stage sampling designs for gene association studies. *Genetic Epidemiology* 2004; 27:401–14.

27 Software for Genomic Data

Robert Gentleman

CONTENTS

27.1	Introduction	489
27.2	Getting Started	491
27.2.1	Data Resources.....	493
27.3	Data Analysis	494
27.3.1	DNA Microarrays.....	494
27.3.1.1	Image Processing	495
27.3.1.2	Normalization	496
27.3.1.3	Quantification of Expression	497
27.3.1.4	Differential Expression	498
27.3.1.5	Software for Expression Analysis.....	499
27.3.2	Other Technologies	500
27.3.2.1	Protein Mass Spectrometry	500
27.3.2.2	Array-CGH	501
27.4	Integrated Databases	503
27.5	Machine Learning	504
27.5.1	Supervised Machine Learning	505
27.5.2	Unsupervised Machine Learning	505
27.5.3	Visualization.....	508
27.5.4	Data Re-use	508
27.5.5	Machine Learning Software.....	509
27.6	Software for Graphs and Networks	509
27.7	Reproducible Research	511
27.8	Software Development.....	511
27.8.1	Data Technologies	513
27.9	Discussion	514
	Glossary.....	514
	References	516

27.1 INTRODUCTION

The relationship between cancer and genetic, or genomic, defects is well established, and in a sense cancer can be characterized as a disease of the genome. Hence high throughput technologies for measuring different aspects of the genome and of the

molecular biology of the cell are clearly of interest to those involved in cancer research. The development of computational tools, models, and statistical methodology to complement the rapidly evolving experimental techniques is essential. In this chapter we consider the computational aspects of this scientific revolution from several different perspectives. These include the problems of finding and assessing software solutions, using software tools, and creating software tools.

In recent years the ability to collect biological data, particularly at a molecular level, has greatly expanded. Genome wide screens are becoming routine and hence very large data sets, collected using different techniques and under different experimental conditions, are becoming common place.¹ Adequately processing the data, performing quality control, and subsequently incorporating the data into a statistical model is a daunting task facing many practitioners. New terminology is proliferating, and the size and complexity of the experimental data are large. Further, making sense of the experimental data usually requires associations with biological meta-data, such as the chromosomal location or sequence information, protein complex comembership, and so on.

The variety of data generating technologies is impressive. DNA microarrays are used to measure the relative abundance of tens of thousands of mRNA species; array comparative genomic hybridization (aCGH) is being routinely used to measure DNA copy number; digital karyotyping² and SAGE³ use restriction-enzyme and sequence-based technologies to provide digital information about mRNA and DNA copies. New arrays for detecting single nucleotide polymorphisms (SNPs) can be used for loss of heterozygosity studies, DNA methylation studies and many single cell assays are being developed and used. In addition there are many other technologies being explored to study the proteome. This includes phosphorylation tools, the yeast two-hybrid system⁴ and affinity purification methods^{5,6} to assess protein–protein interactions, SELDI-TOF, and other time-of flight technologies^{7,8} to estimate relative polypeptide abundances. Each of these very many different technologies has its own set of specific technical problems, and being even moderately conversant with them is time consuming.

While the experimental data are typically large and complex, they are seldom analyzed without reference to the growing body of biological metadata. Information about genes, such as their location, the processes they are involved in, their transcriptional activators, common mutations, and so on is quickly being accumulated. Large databases, sometimes species specific, are being constructed and populated. The use of these metadata can provide substantial meaning to an analysis, and they should be used when possible. However, making use of the metadata requires additional software for obtaining them and for aligning them with the experimental data that are to be analyzed.

Scientific investigators are confronted with large, complex data sets that require substantial preprocessing and modeling. These tasks will be carried out using software, but because many of the tasks are new, that software will be unfamiliar, or in some cases, as yet unwritten. Because the field is immature, those tools that are available tend to be complex, incomplete, and not necessarily easy to use. Unfortunately in such a milieu the tendency is to write small scripts and to carry out one-off analyses. Such an approach cannot be condoned as it is likely to fail on one of the most

important aspects of scientific investigation, that of reproducibility. There are good arguments for making the computational aspects of every analysis completely reproducible. Unlike the experimental situation where lab to lab variation is to be expected, the computational analysis should yield numerically identical results regardless of where it is done, provided exactly the same input data are used and the same transformations and algorithms applied. Reproducibility should be a required standard of any analysis.

In this chapter we consider some of the software tools available for analyzing genomic data, we discuss how to locate software, and finally we give some guidance on strategies for writing and delivering software and data for others. While many investigators will only be interested in software use, there is a considerable fraction of scientists that will carry out some form of software development. In part they will do so because specialized software is often needed for performing the appropriate analysis. In other cases they will have devised a new and improved methodology and will want to share it—in computational biology new ideas are adopted if software is available, and they tend to be ignored if there is no reasonable implementation. Our experience has been that software engineering is a necessary skill for any innovative research in this area.

We note that all software, commercial or open source, comes with a license. That license specifies the conditions of use, and we strongly encourage adherence to the license specified by the author. We also encourage users to appropriately cite the software that they are using. Almost all software projects have a primary publication that can and should be used whenever the software is. For example, users of the R project are asked to use the following citation: Ref. 9 which can be obtained in R via the function `citation`.

The outline of the chapter is as follows. We begin by making some general comments about software and where to find it. We then consider software in the context of several different specific settings but with an emphasis on the analysis of data from DNA microarrays. Next some of the issues that arise in finding and using metadata are discussed; this is followed by an overview of different machine learning problems and some of the available technologies. We then consider notions of reproducibility and end with a discussion of software development. Our emphasis will be on software available from Bioconductor, but we also discuss many other projects and software tools that can be used. R and Bioconductor packages are shown in a sans serif font, e.g. `limma`. A glossary contains references to all software that is mentioned.

27.2 GETTING STARTED

When confronted with genomic data, an investigator will need to obtain appropriate software tools and begin processing the data in order to answer the scientific questions of interest. A wide range of software is available, both commercial and open source. Investigators are encouraged to look quite broadly at the options available, as the offerings tend to be quite different, and whether a tool is useful or not depends substantially on the objectives and capabilities of the user.

In choosing software to use, the best recommendation we can give is to find software that you feel comfortable with and that is capable of addressing the problems that

you will be working on. The more it resembles other tools you are already familiar with, the easier it will be to use. We also suggest that you leave adequate time to familiarize yourself with the tools and the analysis objectives. It is often very difficult to learn how to use new software tools at the same time that you are learning about a new technology. We also caution readers to check to see how old the software is, especially if they are unfamiliar with it and if they have obtained it online. The technologies and computational methodologies are constantly being updated and improved, and older software may not be capable of dealing with new data.

It is also a good idea to see what software others around you use. It is often the case that universities or labs have licenses for suites of software tools, and it is easier to learn how to use software from someone than from a book. You should make sure that the software is capable of fulfilling your needs. Many point-and-click software tools do not have a sufficiently rich set of analysis tools to accommodate an analysis of some of the more interesting experiments. It is easy to find software for carrying out preprocessing and performing two-sample comparisons, but those are about the only really common functionalities. Good resources for finding software tools are original journal articles (although these often do a bad job in citing the tools used), online search engines, domain specific e-mail discussion lists, or news groups. Some journals publish papers based on software tools, and in particular we mention *Bioinformatics*, the *Journal of Statistical Software*, and *Genome Biology*, all of which have sections that are oriented toward software-related papers or to the publication of software itself.

If a graphical user interface (GUI) is desired, then there are a number of commercially available tools such as GeneSpring, Spotfire, or Xpression NTI. Free or open source tools worth considering include TreeView, Cluster, and ScanAlyze from the Eisen lab, dChip, limmaGUI, and BRB-ArrayTools. While GUIs can substantially lower the barrier to entry, they also tend to be limiting because only preprogrammed analyses are easy to carry out. For many programs it is possible to extend the GUI and to add purpose built menu items to carry out relevant analyses, but such extensions can themselves be quite complex.

Command-line applications tend to give more flexibility at the cost of a user interface that tends to be more difficult to master. Bioconductor, Bioperl, and Biopython are among the major projects here. They provide different tools and interfaces, and users should select a language that they are familiar with or want to learn. These projects tend to rely on a distributed development strategy and on substantial user contributions. The software can be of quite different levels of maturity, but it tends to be closer to the leading edge of research than the GUI applications (dChip being a notable exception).

Another advantage of command-line applications is that it is generally possible, one might even say necessary, to construct scripts to carry out an analysis. This facilitates both reproducibility because the same script can be run again sharing because a script can be e-mailed to others for their use.

A good strategy for getting started is to attend a short course on genomic data analysis. There are many given all over the world, generally focusing on a single software analysis tool. In some cases, such as those taught as part of the Bioconductor Project, the lectures and laboratory materials are posted on the internet for self-study.

Attending a short course also gives you the opportunity to meet others who are interested in analyzing genomic data and to discuss your specific analysis with a motivated peer group.

Once the software tools have been selected there are of course many different parameters that must be chosen and methods decided on before the data can be analyzed. There is rather little comparative data available; this is mostly due to the lack of gold-standard data sets on which to do the comparisons. Currently users are largely unable to make truly informed choices, but this will change as the field matures. For Affymetrix data¹⁰ provides some basis for the comparison of different methodologies using a reference data set where different mRNAs were spiked in at known concentrations.

27.2.1 DATA RESOURCES

Experimental data can be obtained from different resources. The Gene Expression Omnibus (GEO),¹¹ from the NCBI is one large source of different gene expression datasets. Other online repositories are ArrayExpress from the EBI¹² and YMD, the Yale microarray database.¹³ In addition many authors provide their original data as supplementary materials that can be obtained either from the journal Web site or from the authors' own Web site.

Few analyses can be completed without making use of available biological meta-data. By biological metadata we mean the large amount of information that is available about the biological system under investigation. The association of the appropriate biological metadata with the observed experimental data is one of the more interesting and challenging aspects of the analysis of genomic data. Two basic strategies are used to access biological metadata. One is to rely on online resources, often using interactive querying tools such as SOAP. The second is to download versions of selected resources and to amalgamate them into a locally curated database. Among the advantages of the first approach are (1) the data are up to date and (2) only those data that are actually used need to be obtained. The disadvantages include reliance on the Internet and data volatility; the major databases update their holdings regularly, and the returned values could change. For locally curated data the advantages include data continuity and the ability to combine several different sources of similar data into a single, more reliable source. The disadvantages include the effort needed to actually carry out the amalgamation and the fact that these data are generally less up to date than corresponding online sources.

The first issue of each volume of the journal *Nucleic Acids Research* has a catalog of public databases relevant to molecular biology. The online category list (www3.oup.co.uk/nar/database/c/) covers nucleotide, RNA, and protein sequence, "structure" databases (small molecules, carbohydrates, nucleic acid, and protein structures), vertebrate and nonvertebrate genome databases, pathways, human genes and diseases, microarray data, and proteomics resources. There are many organism specific database projects, especially for commonly used model organisms, such as Flybase and Wormbase. The National Center for Biotechnology Information (NCBI) is perhaps the most widely known and general source of data. Some of the many sources of meta data are listed below.

- Bioconductor** Provides both prepackaged, curated data as well as tools that allow users to build their own metadata packages and other computational resources (<http://www.bioconductor.org>).
- EBI** A large collection of resources such as nucleotide sequences, peptide sequences, macromolecular structures, microarray experiments (ArrayExpress), genome annotations (Ensembl), and other genomic resources (<http://www.ebi.ac.uk/services>).
- FlyBase** A database of the *Drosophila* genome (<http://flybase.bio.indiana.edu/>).
- GO** The Gene Ontology Consortium (GO), which together with GOA provides both the ontologies and the mappings from genes or gene products to elements of the different ontologies (www.geneontology.org).
- KEGG** The Kyoto Encyclopedia of Genes and Genomes, provides information on pathways, and several other aspects of genomics (www.kegg.org).
- NCBI** An impressive collection of different data resources including PubMed, GenBank, HomoloGene. Many, but not all, of the resources are available through Entrez, a Web services model (<http://www.ncbi.nlm.nih.gov/>).
- NCI** The NCI Center for Bioinformatics contains links to many different resources (<http://ncicb.nci.nih.gov/>).
- NetAffx** The Affymetrix Web site that provides mappings from Affymetrix identifiers to many other identifiers, such as LocusLink, GenBank, etc. (<http://www.affymetrix.com/analysis>).
- SGD** *Saccharomyces* Genome Database SGD is a database of the molecular biology and genetics of *Saccharomyces cerevisiae*.

27.3 DATA ANALYSIS

In this section we consider some of the different specialized data analytic tasks that an investigator might want to carry out. There are many recent books, such as, Refs. 14–17 that cover different aspects of the handling and analysis of genomic data. These books are good starting places to find out about software and appropriate analysis methods for different data types. In this section we discuss some of the most important topics but refer the reader to the references listed above as sources of more detailed information.

27.3.1 DNA MICROARRAYS

By far the dominant high throughput genomic data experiment is the DNA microarray. We now consider some of the software issues and projects that are related to the manipulation and analysis of microarrays. We distinguish between spotted arrays (where the arrays are spotted using some form of printing, regardless of what is spotted on them) and the manufactured arrays (of any technology, such as photolithography or ink-jet technology). The Website: www.statsci.org has a nice section on microarrays that covers many of the important technologies and problems, together with links to software and other online resources.

The general design of such studies begins with the decision about what samples are of interest and the selection of a microarray to use. Subsequently samples are prepared (labeled, etc.) according to the protocol appropriate for the technology being used and

hybridized to the microarrays. Typically mRNA molecules in the samples are labeled with some fluorescent dye, and it is the intensity of the fluorescence that is used to determine the presence and relative abundance of the associated mRNA.

In this setting, data processing begins with the capture of the image using a scanner. This procedure produces an image file, often as a TIFF or similar format. The image file is processed to locate the spots, determine how large they are, and often to separate foreground (the signal we are interested in) from background (the signal we are not interested in). At this point the spots on the array have been found, and some estimate of the intensity of the signal at those points has been made. A number of different software tools can be used to carry out these analyses. GenePix is among the more widely used. For Affymetrix GeneChips, the manufacturer provides a scanner and software that scans the GeneChips to produce `.dat` files; these are then processed to produce `.cel` files. The `.cel` files contain per spot intensities that are further processed to yield per gene (or per probe-set) estimates of mRNA expression levels.

The next step in the processing is generally to *normalize* a batch of arrays so that they are comparable. The purpose of normalization is to ensure that the differences between samples are likely to reflect biological differences rather than experimental or laboratory handling artifacts. In some cases, researchers are interested in determining whether or not a gene is expressed in their samples. Not all genes are expressed in all tissues, and a widely held belief is that approximately 40% of the genome is expressed in any specific cell. Note that this observation implies that in most settings 60% of the measurements are not relevant to the question of interest. Only those genes that are expressed under some set of experimental conditions can provide information about the scientific question that was asked, and excluding probes for genes that are not expressed under any of the conditions should be beneficial. Once this has been done, one of the possible next steps is to address the question of determining which genes are differentially expressed under the experimental conditions. We are using differential expression as an exemplar and note that there are very many other analyses that could be carried out. By differentially expressed we mean that the mRNA is expressed at one level under one set of conditions and at a different level under a different set of conditions. The term *conditions* is interpreted quite loosely and often refers to phenotype. In the next few subsections we consider these steps in more detail and provide explicit references to software that can be used.

27.3.1.1 Image Processing

Once the image file has been obtained, the data analytic processing begins. We must note, however, that image processing has received rather little attention, and much more research has been done on the topics that are downstream from that. The actual tools that will be used depend on the microarray technology that was employed. For spotted arrays commonly used tools are SPOT (<http://www.cmis.csiro.au/IAP/>) and GenePix from Axon Instruments. The UCSF Spot system is another alternative; written entirely in C for Microsoft Windows platforms Ref.18 See either Ref.16 for a general overview of the issues and problems, or Ref.19 for a discussion more focused on cDNA arrays. For Affymetrix GeneChips, vendor supplied software is typically used to carry out all aspects of the sample preparation and processing up to the production of the `.cel` files.

As noted above, the first step in processing the image data is *addressing*, or locating, the spots in the image. Next the physical size of the spot (or foreground) is determined. At the same time the background intensity is also quantified. The background is the intensity that would obtain if there was no foreground. The background can vary quite substantially across the slide, and it is important that adjustments be made for it. The spot intensities are then computed as some functional (sum, average, 75th percentile) of the pixels that are deemed to be part of the spot, typically with some adjustment for the local background. It should be possible to provide some form of standard error on this quantity, although that is seldom done, and even less often is that information used in downstream analyses of the data. However, spot intensity is measured with error, and there are good reasons to consider errors-in-variables models (e.g. Ref. 20)

27.3.1.2 Normalization

Normalization is a procedure that is applied to the experimental units (microarrays in this case) to make them comparable prior to any downstream analysis such as determination of differential expression or machine learning. The methodology used tends to be similar although not always identical across microarray technologies. One of the basic assumptions of most normalization algorithms for microarrays is that among the samples being processed the levels of mRNA expression are relatively constant across samples for almost all mRNAs, or in other words, that there are relatively few differentially expressed genes. This has profound implications, and in experiments where the effect of the experimental protocol is to substantially alter the expression levels of all genes (as might happen with some universal inhibitors of translation or transcription) or that involve different tissues, the method used for normalization should be considered carefully. In such cases new methods for normalization may need to be explored and developed.

A number of different normalization techniques have been used including linear, affine-linear, local polynomials, or nonparametric rank-based methods, such as quantile normalization. Quantile normalization relies on the assumption that the distribution of mRNA abundance is similar across arrays. For each array the measured intensities are ordered from smallest to largest and then the average quantile, averaging across arrays, is computed. Then for each array the ordered expression values are replaced by the average, so once the data have been normalized all sample quantiles are identical across arrays. However, spots correspond to different quantiles on different arrays, and so the estimated gene expression values differ across arrays.

For some types of arrays there are likely to be systematic differences that should be controlled for, in addition to the between array normalization. For example, with spotted arrays the spotting is typically done with between 4 and 16 print-tips, and the quality of the spot can vary substantially across print-tip. Some form of normalization that accounts for differences between print-tips should be considered. Chapter 6 of Ref. 17 and the cDNA chapter of Hsu et al. in this volume deal specifically with these topics.

Normalization is generally essential since there are often systematic differences between the microarrays that reflect differences in experimental conditions, handling of samples, labeling effects, or batch effects that have little to do with the biological questions. If any such effect becomes confounded with a comparison of interest, say

one type of sample was done one week and a second type of sample was processed on a different week, then it may be very easy to distinguish the two types of samples on the prenormalized data. However, this would not provide any biologically meaningful information and would simply reflect the differences in handling and processing. In order to make biologically meaningful comparisons, it is essential that all uninteresting between-sample differences be removed prior to an analysis.

27.3.1.3 Quantification of Expression

In this section we separate the discussion into two parts, one for Affymetrix-like arrays and the other for spotted arrays, as this is one of the places where there are substantial differences.

On Affymetrix GeneChips, each gene is represented by one or more probe sets, and a probe set is some number (generally in the 10 to 20 range) of probes that are specific to that gene. A probe is a sequence of 25 bases, often referred to as a 25mer. There are fairly large differences in the way in which the probe level data are combined to provide probe or gene level estimates. Users may use vendor supplied software (currently MAS 5.0) or software from other parties such as dChip, or either RMA or GCRMA from the affy package 21, 22. A more comprehensive listing of available software is given in Section 5.

Cope et al. have developed a protocol and maintain a Web page dedicated to the comparison of different methods on the analysis of Affymetrix data. Currently methodologies are compared based only on the analysis of a specific set of experiments where known concentrations of a small number of mRNAs have been added. These are commonly referred to as spike-in experiments. Interested parties can use this resource to compare candidate methodologies and select one whose properties are most suitable to their situation.

With spotted arrays a large number of reference sequences are spotted on to an array. Subsequently two different samples of nucleic acid (typically a reference and a test) are labeled (differently) and hybridized to the array. The goal of the subsequent processing is to provide accurate estimates of the ratios of relative abundance of the nucleic acid in the hybridized samples. This is done by processing the image, detecting the foreground (the spots) intensity, and extracting from that the background intensity. Unlike Affymetrix arrays, each nucleic acid is represented by a single spot (although there may be replication), and hence an estimate of the per spot intensity is a direct output of the image processing software.

If the two samples have been hybridized to the array, then it is common to report their ratio. One of the effects of taking ratios on a per spot basis is that much of the background variation is removed because it is generally the same for both samples. In other cases only a single sample is hybridized, and then the analysis is more similar to that for Affymetrix data.

There remain many different data analytic steps that should be carried out prior to any analysis of the data for differential expression. These steps include the creation and use of diagnostic plots, normalization, and different quality control diagnostics. Diagnostic plots often include an examination of intensity as a function of the print-tip used to spot the array. A second diagnostic involves examining the intensity as a

function of the plate that stored the cDNA samples prior to spotting. It is possible that a plate was handled poorly and perhaps dried out. If this happens then there will be no cDNA fixed to the plate, and the labeled samples have nothing to hybridize to. This possibility should be explored and any problems remedied. Other diagnostics involve examining the images for unusual patterns, scratches, or other defects. Virtually any high level software with tools for exploratory data analysis, such as box plots or cumulative distribution functions, can be used; see the chapter by Hsu et al. in this volume for examples.

Tools for carrying out these different procedures often accompany the image processing software (e.g., GenePix), or they can be obtained by using software such as the *marray* package²³ or the *vsu* package.²⁴

27.3.1.4 Differential Expression

Determining which genes are differentially expressed under the different experimental conditions is of primary interest in analyzing DNA microarray experiments. The experimental setting provides substantial difficulties because the number of experimental units is typically one or more orders of magnitude smaller than the number of covariates. Many different strategies are employed in attempting to address the problem, and most are heavily reliant on software. This subject has been widely discussed; see, for example Ref. 25 or 26. The Bioconductor project has very many different software tools that have been developed for analyzing such data; of particular note are *limma*, *EBarrays*, and *genefilter*.

We begin by emphasizing that the current state of the technologies involved permits comparisons between samples for a fixed probe (or gene), but it does not permit comparisons within samples between genes. The problem is that while the observed intensity at a probe is proportional to the abundance of the complementary mRNA, a number of factors prevent an estimate of the absolute abundance. Hence, it is valid to consider the probe as fixed and compare across samples. This is true for both two-color arrays and single channel arrays.

It is also worthwhile remembering that the estimated expression values do not correspond to a single cell but rather to an average over many thousands of cells. If the population of cells assayed is diverse, then the estimated expression values will reflect that diversity. Recent advances such as using laser dissection to better purify the samples can improve the situation, but these techniques introduce their own biases and problems.

While the amount of mRNA required for an analysis seems quite small, it is not always possible to obtain a sufficient quantity, and in those cases some form of amplification is often carried out. Such procedures are unlikely to affect all mRNA species equally. Hence, relative abundance of mRNAs within samples after amplification may have changed quite substantially, further impairing the validity of any within-sample comparisons. However, it does seem reasonable to assume that any effects are based on mRNA sequence and hence will be relatively constant across samples, so that the within-gene between-sample comparisons should remain valid. We must emphasize that it is essential that any special treatment (such as amplification, laser-capture microdissection) be applied to all samples. If the samples are treated differently, the ability to compare between samples can be lost and along with it all hope of valid inference.

Finally, we recognize that while the bulk of the research and analyses of these data has focused quite narrowly on the selection of sets of differentially expressed genes, such a view is quite limited. The assays have taken a global picture of gene expression, and because gene expression is typically carefully orchestrated by different cellular processes, it seems very likely that useful information will come from a more holistic analysis of the data. For example Ref. 27 suggest that it is of more interest to identify pathways whose members all change by a small but similar amount than to identify genes with large changes in expression. Detecting such changes is essentially impossible with the rather prevalent gene-at-a-time approach that dominates current practice.

However, gene-at-a-time methods are popular, and there are some general concerns that should be addressed. Among the more extensively researched methodologies are the p -value correction methods, and large amounts of journal space have been devoted to this topic. There are a number of software solutions available for almost all recommended approaches (e.g., the R package `multtest`). However, p -value correction methods are not a panacea, and their effect is more akin to that of a band-aid than a solution to the problem of detecting important biological differences. By and large, p -value correction methods work by adjusting the level of the p -value deemed to be interesting downward. And by lowering this, the set of genes deemed interesting gets smaller and tends to be enriched for genes that are biologically interesting. But this comes at a cost. As the level is lowered, some genes that are truly differentially expressed will be excluded. The only real solution seems to be to bring more biological data to bear on the problem, to reduce the set of genes tested, and to test a few very specific hypotheses before engaging in exploratory data analysis.

A popular approach to identify differentially expressed genes is to conduct a standard statistical test, such as the t -test, for each gene separately. Kendzierski et al.²⁸ note that treating genes as separate, independent experiments tends to be less efficient than adopting a Bayesian approach and taking advantage of the shared information between genes, Baldi and Long,²⁹ Tusher et al.³⁰, Lönnstedt and Speed,³¹ and Smyth²⁶ all proposed moderated versions of the t -statistic where the denominator is the sum of the usual gene-specific variance estimator and a constant estimated using all of the data. When there are relatively few samples available the gene-specific variance estimates tend to be quite variable and this approach provides some stability.

While a testing paradigm can be usefully employed, most biologists also are interested in the *fold-change*, which in the two-sample case is simply the ratio of the two means. In other settings, such as more complex designs, analogs of the fold-change can be computed using ratios of estimated effects. Irrespective of the number of replicates, the use of a fold-change criterion may be beneficial as this helps to screen out genes whose effect sizes are small in absolute terms, even though they may be statistically significant.

27.3.1.5 Software for Expression Analysis

Here we list a few of the very many software projects that are producing and distributing software tools appropriate for different analyses of gene expression data. Because this is a rapidly changing field, a good strategy is to make use of your favorite Internet search engine.

affy Bioconductor software for a complete and comprehensive treatment of Affymetrix data from CEL files to expression estimates. A wide variety of normalization methods and methods for estimating gene expression are available.²¹

BRB-ArrayTools A Windows-based set of microarray analysis tools. It is free for noncommercial use and is available at <http://linus.nci.nih.gov/BRB-ArrayTools.html>. They also have regularly scheduled training sessions. Some of the functionality is discussed in Ref. 17.

Cluster Is available from the Eisen lab (<http://rana.lbl.gov/>), together with TreeView, ScanAlyze, and other software. The software is free for non-commercial use and is widely used.

dChip dChip³² provides a Windows-based analysis tool for Affymetrix microarray data. The input is from .CEL files, and the software provides expression estimates, quality control, visualization, and data analytic capabilities.

EBarrays An R package that takes an empirical Bayes approach to determining differential expression,³³ available through Bioconductor.

limma Software for fitting linear models to expression data. This package has a nice graphical user interface provided by the limmaGUI package, and it is available through Bioconductor.²⁶

MAANOVA Software in either Matlab or R from the Churchill lab at the Jackson Laboratories.³⁴

marray Bioconductor software for the processing of spotted arrays. The inputs are a wide variety of vendor specific output formats. The presumption is that the image files have been processed and different quantities such as foreground and background for the different channels are available. The outputs are estimated red–green intensities, typically on the log scale.²³

MAS 5.0 Affymetrix supplied software for estimates of gene expression for their arrays.

TM4 Consists of four major applications, Microarray Data Manager (MADAM), TIGR Spotfinder, Microarray Data Analysis System (MIDAS), and Multiexperiment Viewer (MeV), as well as a Minimal Information About a Microarray Experiment (MIAME)-compliant MySQL database. All are freely available from TIGR and are produced by the Quackenbush lab.

vsn Bioconductor software for variance stabilizing normalization.²⁴

On the commercial front there are very many different options including MATLAB with a new bioinformatics toolkit and the Insightful Corporation that markets the ArrayAnalyzer suite of tools for S-Plus, which is largely based on Bioconductor materials.

27.3.2 OTHER TECHNOLOGIES

27.3.2.1 Protein Mass Spectrometry

A number of different high-throughput protein mass spectrometry technologies are becoming widely used. One type of mass spectrometry makes use of the time it takes

a charged particle to transit a drift region, usually a tube, in order to assess the molecular mass of the particle. Among the more popular time-of-flight technologies are MALDI-TOF and SELDI-TOF. We will consider SELDI-TOF (since that is where our experience is). SELDI-TOFMS is *surface enhanced laser desorption/ionization time-of-flight mass spectrometry*. It is a technology for detecting and measuring proteins from tissue or body fluid samples. Papers on this topic are starting to appear; those of note include Ref. 8, 35, and also Yasui et al. in this volume. It is hoped that this technology will lead to the discovery of important biomarkers for cancer; see for example Ref. 36, but also Ref. 8 for a different interpretation of the Petricoin data.

The methodology is typically applied to serum or other tissue, and a common situation is the comparison of two or more groups. Each sample produces a spectrum such as that shown in [Figure 27.1](#), in which peaks correspond to polypeptides, and it is believed that the height of the peak reflects the abundance of that polypeptide in the sample. The horizontal axis corresponds to m/z , which is mass over charge, while the vertical axis corresponds to a measure of abundance.

Many statistical problems must be dealt with including:

- Baseline correction: the process of adjusting the spectra so that the baseline is zero
- Cut-off selection: selecting the m/z value that corresponds to the point where the effect of the matrix noise becomes less substantial
- Quality assessment: ensuring sufficient signal in the observed spectrum to allow for further analysis
- Peak detection: the identification of those m/z values that correspond to peaks in the spectrum
- Addressing or peak alignment: potential for a small amount of drift in the horizontal axis; alignment carried out subsequent to peak detection
- Normalization: attempt to make all spectra comparable and to thereby ensure that detected differences are not artifacts of the sample collection or handling

The final result is a set of baseline subtracted, aligned peaks. See [Figure 27.2](#) for the processed version of the spectrum shown in [Figure 27.1](#) together with the peaks that were detected. For SELDI-TOF data software is available from Ciphergen (the company that manufactures the system) and from the Bioconductor Project in the form of a software package called **PROcess** largely authored by X. Li.

27.3.2.2 Array-CGH

Microarray-based comparative genomic hybridization (aCGH) is a technology that measures DNA copy-number, which can be mapped directly onto the genomic sequence and hence allows us to find regions of loss or gain of DNA in somatic cells. Many different microarray technologies can be used to hybridize the genomic DNA. The output is typically the ratio of copy number between two jointly hybridized samples (and hence it is like the two color spotted arrays). But because one can generally find genomic DNA with the appropriate number of single sets of chromosomes

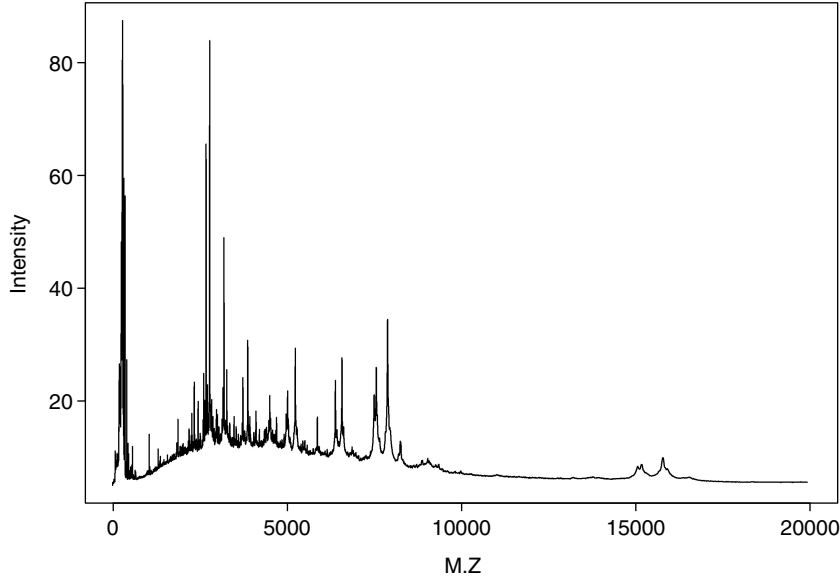


FIGURE 27.1 Raw spectrum of polypeptide intensity by mass over charge (m/z) from a serum sample. The main heading is the sample identifier.

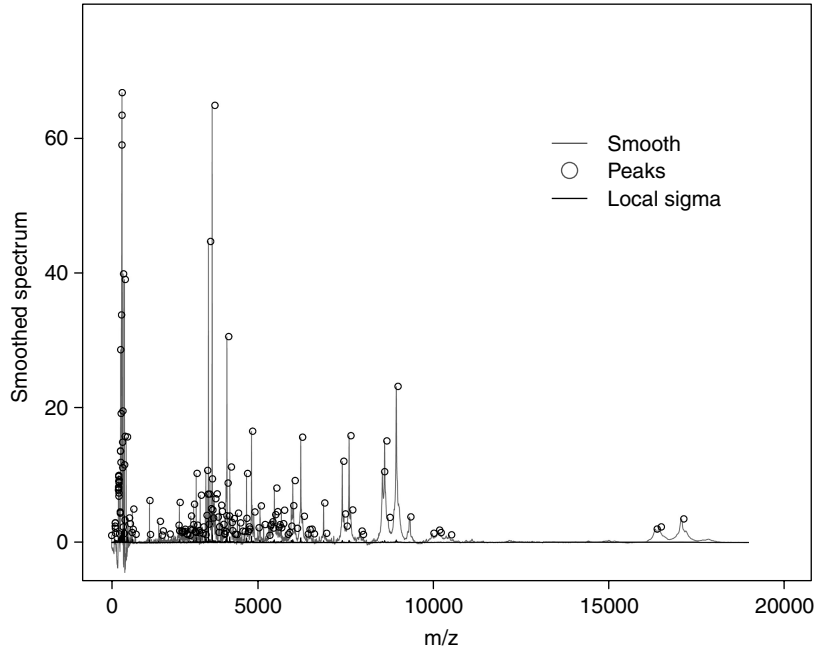


FIGURE 27.2 Visualization of the spectrum from Figure 27.1 after baseline subtraction, addressing, and normalization have been carried out. The smoothed spectrum, estimated peak locations, and local variation are displayed.

(i.e., *ploidy*), the outputs can usually be interpreted in that context. See Ref. 37 for a description of one type of array.

Many of the issues that concern microarrays must also be addressed for array CGH data, such as image processing, background subtraction, and so on; see the chapter by Hsu et al. in this volume. Often the same tools are used (and in many cases the same arrays). Once the data have been processed and basic quality control procedures have been carried out, the next step in the analysis is the assignment of copy-number values to different regions of the genome for each sample. A number of different methods are commonly used. One is based on hidden Markov models, and a second is based on a circular binary segmentation algorithm. These methodologies are available in two R packages, **aCGH** and **DNACopy**, both available through the Bioconductor Project.

27.4 INTEGRATED DATABASES

The proliferation of high throughput experiments and the obvious benefits of carrying out integrated analyses have led to the creation of a number of tools. Notable among them are ONCOMINE,³⁸ SOURCE,³⁹ BASE,⁴⁰ and caCORE.⁴¹

In fairly generic terms these four projects are quite similar, although in practice they are very different and clearly aimed at different sets of users. The basic set of problems that they need to address is :

- They must define and implement a database schema that is appropriate for the type of experimental data they will use (typically this has been microarray data).
- They must create and curate the biological metadata that they have decided to use. This is then associated with the probes (or biological components) of the experimental data. Further, they must develop some framework for describing the experiments that are available and ideally provide some indication of the design. If patient samples are used, then covariate data on these will be needed.
- They must create and implement a querying interface. Typically it supports selection of genes (either by name, or through some aspect of the biological metadata such as pathway or function). They also provide tools for selecting appropriate experiments.
- Finally, given that the user has selected a set of genes and a set of experiments, they then typically provide some data analytic tools for both visualizing the data and for testing a small set of hypotheses.

It should be noted that there remain substantial problems, and these tools probably should be viewed as initial efforts rather than as refined and mature tools. Among the problems are the lack of good semantics for describing the different types of experiments that might be available. Time course experiments, designed experiments, and cohort studies are all common but require substantially different methods of analysis. These problems are often side-stepped by the project deciding to focus on only one type of data; for example, ONCOMINE tends to concentrate on two group comparisons.

Further, it is not completely clear that one can or should blindly combine experimental data. Experience with cancer clinical trials suggests that this may not be an entirely appropriate approach. Rather, aspects of meta-analysis are likely to be appropriate, and future generations of these projects are likely to include more sophisticated tools such as those described in Ref. 42–43. The first of these technologies is available as an R package, *MergeMaid* through Bioconductor.

As noted above, the merits of combining large genomic data sets to provide a more comprehensive view of coordinated changes at the genomic level is clearly appealing. These databases and similar projects will provide resources for data mining such as finding genes whose patterns of expression exhibit similar changes across cancer types. The information contained in several experiments should provide stronger evidence than single experiments. Further, in many cases, it should be possible to test hypotheses generated based on one experiment with respect to similar existent data.

27.5 MACHINE LEARNING

Once all of the data have been processed and assembled, the next task is generally some form of statistical analysis or machine learning. In essence machine learning is a specific subset of statistics; however, machine learning has the advantage of having a well defined vocabulary that is used consistently, and for that reason alone it is worthwhile adopting their terminology. Ideas such as model fitting or exploratory data analysis are not generally part of machine learning, although Hastie et al.⁴⁴ do take a more general view. In statistics the concepts of clustering and classification are hopelessly muddled, with different sets of authors using the same terminology to describe completely different ideas. So in this section we define and then use some standard terminology from the machine learning community.

Machine learning falls into three general categories: unsupervised learning (usually referred to as clustering in the statistical literature), supervised machine learning (usually referred to as classification in the statistical literature), and feature selection (conceptually close to notions of variable selection in regression models but a little more general and in bioinformatics often referred to as finding signatures of gene expression). Only the first two of these are covered in the subsequent discussion. We augment the discussion somewhat by including a brief discussion of sample reuse methods such as cross-validation, the bootstrap, and permutation methods.

There are very many good books available on the topic of machine learning, in particular, Ref. 45–47 are all very valuable resources. There are good implementations of most algorithms in many different languages and readers are encouraged not to reimplement these routines but rather to make use of existing implementations.

Both supervised and unsupervised machine learning methods rely on a definition of distance between samples and in some cases between sets of samples. In many software implementations, the distance used is sometimes hidden and cannot be explicitly set by the user. It plays an essential role nonetheless, and practitioners should pay careful attention to selecting an appropriate distance and to ensuring that the same distance is used in all machine learning methods. This can often be achieved by pretransforming the data so that Euclidean distance on the transformed data is equivalent to the desired distance on the untransformed data. Almost all machine learning algorithms use or allow the selection of Euclidean distance, so this technique tends to be quite general.

27.5.1 SUPERVISED MACHINE LEARNING

Supervised machine learning begins with data for which the variable of interest, the outcome, and the predictors, or features, are all known. The goal is to build a model, using the features, that can be used to predict the outcome of new samples, for which only the features are known. The problem can be described quite generally in terms of decision theory, and while most software implementations provide an estimated class for every sample that is presented to them, it is possible to augment the set of predictions with two additional categories. It should be possible for a classifier to express *doubt*, that is, a situation where the new sample could reasonably come from more than one class. In other cases, classifiers should be able to report that a new sample is an outlier, in the sense that its features are sufficiently different from those that were used to build the models that it is unlikely that the model applies.

Essential ingredients in building a classifier are the set of features that will be used and the choice of distance measure that is used to say when two samples are similar or not similar. It is well known that building a classifier on the available training data tends to overfit and to provide a biased estimate of how well the data can be classified. In most cases either a test set or cross-validation is used to assess the performance of the classifier. It is generally better to use cross-validation for this purpose.

Tusher et al.³⁰ proposed a number of methods such as prediction analysis of microarrays (PAM), which carries out sample classification from gene expression data by the method of nearest shrunken centroids. Software is available as the R package `pamr`. This methodology has no relationship to the preexisting partitioning around medoids (PAM) method that is described below.

27.5.2 UNSUPERVISED MACHINE LEARNING

In unsupervised machine learning, only measured features are available; there is no outcome, or class, that should be predicted. Sometimes this process is referred to as *class discovery*. It is important to remember that any two objects are different in infinitely many ways (this is basically the ugly duckling theorem), and the hope of categorizing samples into meaningful subsets must rely on domain specific knowledge. For example, when considering gene expression data, a common question is “How many groups are there in my data?” But the answer to that question depends on which genes you would like to consider. The answer may be quite different if only genes in the apoptosis pathway are considered or if only those with known tyrosine kinase activity are used. There can be no simple answer, and domain specific knowledge is essential.

There are a wide variety of clustering methods in use. One general dichotomy includes the hierarchical methods and the partitioning methods. Hierarchical methods provide a hierarchy of nested clusters from one to the number of samples. Partitioning methods divide the samples into some number of prespecified groups, or clusters, and then basically swap members from one group to another to optimize some criterion, which is generally based on having objects within a cluster be similar while the distance between clusters should be large. Because there are many definitions of distance and many ways to define a distance both within clusters and between clusters, the result is a very large number of techniques. In all but a relatively few cases, the implementations are not deterministic since the computation

involved is prohibitive. So different starting values should be used and the outputs compared because they often result in different clusters.

The output of hierarchical clustering is often presented graphically as a dendrogram. It is important to be aware of the weaknesses of this representation. First, and most importantly, it is *not* a visualization method. Visualization is the process of revealing structure that exists in data. Hierarchical clustering, and hence the dendrogram, imposes structure—whether it is real or not. As we have noted above, some notion of between-observation distance must have been used to compute the hierarchical clustering, and so one can ask whether the dendrogram faithfully represents this original, and ideally carefully selected, distance. The cophenetic correlation measures the relationship between the original between-observation distances and those imposed by the clustering. If that correlation is high, then the hierarchical clustering is a reasonable approximation to the original data.

The hierarchical methods can be divided into two subclasses, the agglomerative methods, which start with each observation in its own group and combine groups until there is only one group, and the divisive methods, which start with all observations in one group and repeatedly divide the observations until each is in its own group. Agglomerative methods are the only methods that can be easily made deterministic, and that may be the reason for their widespread popularity; however, they

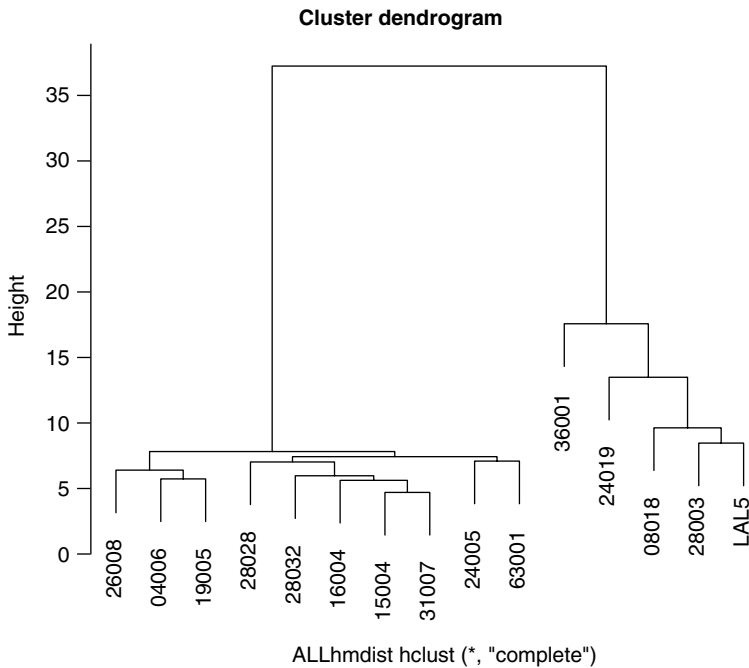


FIGURE 27.3 Dendrogram of gene expression data for leukemia patients. Patient samples were clustered according to a set of relevant genes. This is essentially the same dendrogram as found in [Figure 27.4](#).

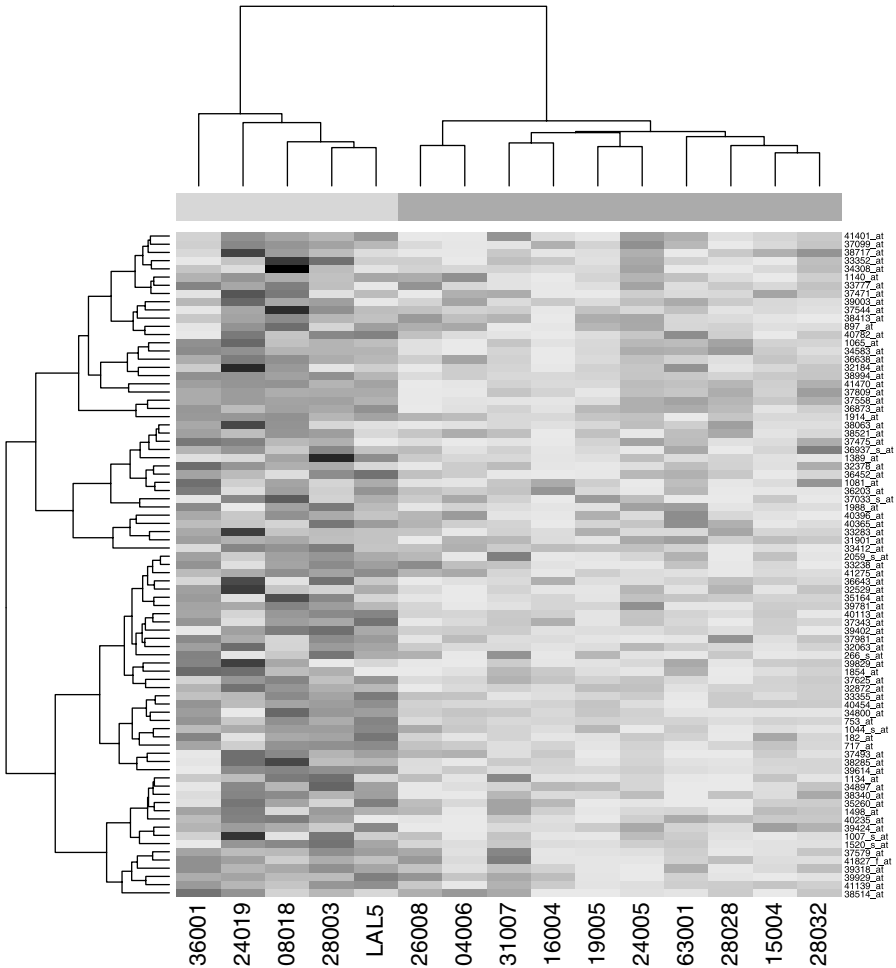


FIGURE 27.4 (See color insert following page 458.) A heatmap comparing leukemia patients. The patients are the columns; the genes they are clustered on are the rows. Shading indicates relative gene expression.

can be criticized on many grounds. The algorithm used is greedy, and once two observations have been joined they cannot later be separated. Hence, it seems likely that such methods provide a more accurate representation of the leaves of the resulting dendrogram than of the upper portions but it is the upper portions that are often used to find a small number of groups. It is difficult to look at [Figure 27.3](#) and not get the impression that there are two groups. The point is that our visual focus is primarily on the upper part of the dendrogram, which for agglomerative clustering may be the least well determined part. Using divisive hierarchical clustering may yield better results when the goal is the identification of a small number of groups.

For partitioning methods, both k-means and partitioning around medoids (PAM) are popular. The former is available in R in the `class` package, while PAM is implemented as `pam` in the `cluster` package. Many other software packages provide implementations of these methods.

Self-organizing maps⁴⁸ (SOMs) are another popular choice of method for unsupervised learning. They are sometimes described as if they are a supervised learning method, but they are not. To begin, clusters are arranged in a two-dimensional grid and samples assigned to clusters. Then a series of iterations that perturb both the cluster centers and cluster membership are applied. The amount of change is controlled by a mixing parameter whose value gradually reduces. Upon convergence the output is a set of clusters and their locations in the plane. Different software implementations are available, many such as the R package `som` are based on SOM PAK (<http://www.cis.hut.fi/research/som>).

27.5.3 VISUALIZATION

Visualizing clustered data has not received large amounts of attention, although there are some tools available. Among the available tools are the silhouette plot⁴⁹, which can be applied in most cases; the use of multidimensional scaling has also been relatively widespread. In Ref. 50 the *heatmap* was introduced as a method of displaying gene expression data. It is a nice display for such data because the rows and the columns of the display can be sorted independently. Rectangular regions of relatively constant color distinguish sets of samples and genes that have relatively constant expression values. This method is widely used, and an example is provided in Figure 27.4.

27.5.4 DATA RE-USE

Because the size of the experiments, in terms of number of samples, tends to be quite small, a number of data reuse techniques are often employed. These include cross-validation, bootstrapping, permutations, and simulations. Given the size of the data sets involved and the complexity of the models, these tools can add a substantial computational burden. However, because there is often little theoretical justification for a specific analysis, and even when there is, some of the important assumptions will be seen to be violated, these data reuse strategies provide a valuable basis for inference.

Cross-validation is basically the procedure of dividing the available data into different sets. These sets usually form a partition of the data, and commonly used numbers of partitions are ten and n , the number of observations. A model is fit using one portion of the data, and the error rate, or misclassification rate, is estimated using the portion of the data held back, which can be as small as one sample. The procedure is repeated until all observations have been used in both ways and the results averaged. It is worth noting that there is no reason that the data must be partitioned (although partitions are likely to be more efficient in the statistical sense) and one can simply generate a large number of splits of the data into two groups.

Cross-validation can be used for model selection or for the selection of parameters to be used in a model. For example, if k nearest neighbors is the chosen classification

method, then an optimal value for k can be selected by estimating the cross-validation error rate for a number of different values of k and selecting as the one to use that value that corresponds to the minimum of the estimated cross-validation error rates.

Bootstrapping is a similar method; here a sample of size n is drawn with replacement from the observed data. A model is fit to the bootstrap sample; typically information is aggregated across all bootstrap samples. Those observations that were not selected can be used (they are often referred to as the *out of bag* sample) to assess the classification error rate. One must be somewhat careful in using bootstrapping for machine learning problems because virtually every bootstrap sample has some exact repeats and many machine learning algorithms behave peculiarly in that setting.

Permutation testing is also widely used, although the implications are not always clearly stated. In most cases one can test for an association between the labels on the observations (e.g., tumor versus normal) by permuting these labels.

27.5.5 MACHINE LEARNING SOFTWARE

A list of some of the different software resources that are available as R and Bioconductor packages for machine learning is given next. However, it is incomplete, and many more tools are being developed. MatLab and other programming languages provide a range of methods as well.

class Software implementations of a large number of machine learning algorithms

cluster Software implementing the techniques described in Ref. 49.

e1071 A package with a number of machine learning algorithms, in particular an implementation of support vector machines.

gpls An implementation of generalized partial least squares, as described by Ding and Gentleman.

MASS The software complements to Ref. 45, which include a rich set of tools for statistical analysis and machine learning.

randomForest An implementation of Breiman's⁵² random forests technology.

27.6 SOFTWARE FOR GRAPHS AND NETWORKS

In this section we consider some of the software tools available for dealing with graphs and networks. These are mathematical abstractions that can be used to describe a set of objects (the nodes) and the different binary relationships between those objects, represented by the edges. Graphs and networks arise in many different contexts in computational biology and bioinformatics. They are a natural data structure for describing relationships between entities such as the bipartite graph that describes the relationship between proteins and protein complexes. Graphs are used in different clustering algorithms;⁵³ they can be used to test for relationships between different biological behaviors^{54,55} and they can be used to describe pathways, to name but a few of the uses graphs have already been put to.

Software for working with graphs can be broadly grouped into three different areas, although there is often overlap between the different uses. There is software that is used

to represent the graphs, that is, to provide the necessary data structures. There is software that implements different graph algorithms such as LEDA⁵⁶ and the Boost graph library,⁵⁷ and finally there are software tools to help in laying out graphs such as Graphviz,⁵⁸ Tulip⁵⁹ or Pajek (<http://vlado.fmf.uni-lj.si/pub/networks/pajek/>).

Many of these tools either already support integration with other software or are moving in that direction. As part of the Bioconductor Project interfaces to Boost (RBGL) and Graphviz, (Rgraphviz) have been developed, in conjunction with the `graph` package that provides a number of different implementations of the necessary data structures. This represents a unique juxtaposition of graph software in a data analytic environment and can be quite helpful for analyzing data.

Pathways, such as metabolic pathways or signaling pathways, are often represented as graphs and networks. As noted previously KEGG provides data on many different pathways, although the current representation is not especially amenable to computation because it is difficult to determine the different relationships between nodes in the graph (and hence the placement of the edges) in a programmatic fashion. One of the pathways from KEGG rendered by Rgraphviz is shown in Figure 27.5.

A recent development is the cMAP project from the NCICB. This project provides us with a set of tools for programmatically manipulating pathways. Pathways have been obtained from both BioCarta (<http://www.biocarta.com>) and from KEGG,

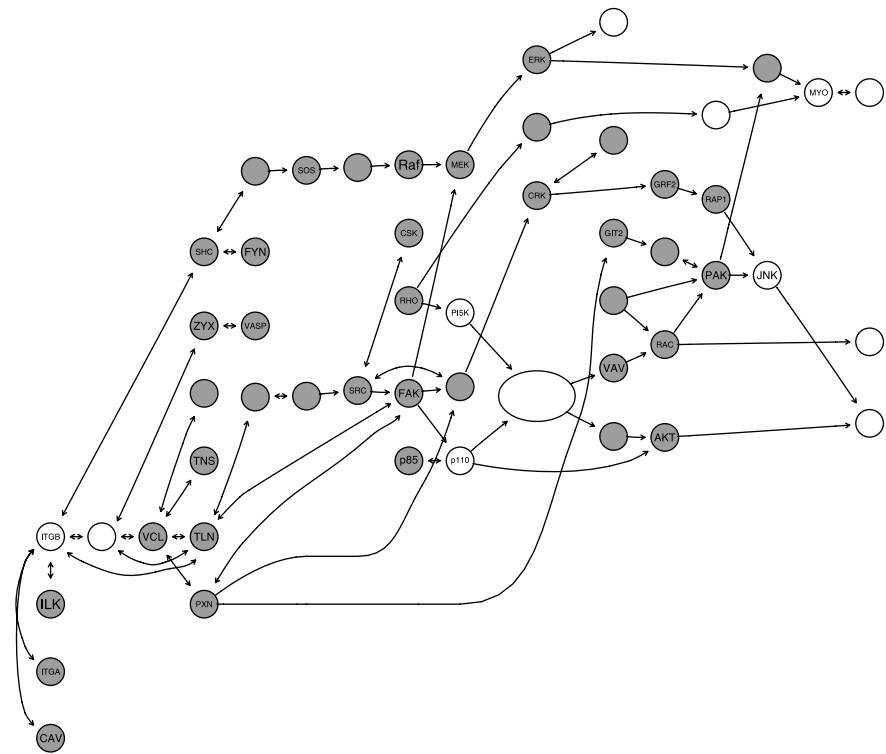


FIGURE 27.5 Integrin mediated cell adhesion network.

and the data have been reorganized to allow for relatively easy access. The two different sources are kept separate, and there is one set of mappings given for KEGG and a separate set for BioCarta.

27.7 REPRODUCIBLE RESEARCH

It is essential for this field to begin to address the important problem of being able to reproduce the results reported by others. We believe that an important goal is to ensure that others can easily reproduce the tables, figures, and important facts that are reported in any scientific paper. While this seems like a daunting task it is not actually that difficult, and there are a number of papers that directly address these issues. Examples include the notions of literate data analysis,⁶⁰ and reproducible research.^{61,62} The tools described in these papers are largely based on R and the Sweave system,⁶³ but they could easily be extended to other languages and word processing systems, as there is nothing that is particularly centric to R.

By *reproducible research* we do not mean any sort of independent verification, but rather that a reader should be able to start with the same data that the authors did, apply the transformations, and carry out the statistical modeling that was described in the paper and thereby draw the same conclusions. It is quite amazing to discover how few scientific papers actually meet this rather straightforward criterion. In part this is due to a lack of suitable tools and paradigms. For example, there is a widely held belief that one need only provide raw data together with a set of scripts that were reported to carry out the analysis. However, this strategy has simply not worked. Often some small changes made to the data, external to any of the usual processing methods, are missed. And once the *audit trail* has been broken, the outputs can no longer be produced from the inputs. See Ref. 8 for one of the more spectacular examples of the problems experienced when attempting to reproduce an analysis.

27.8 SOFTWARE DEVELOPMENT

No matter what experimental data you encounter, you are very likely to have to write some software, possibly not much, in order to analyze them. To do so you will need to make some decisions: what language to use, whether the software should be designed for others to use, whether or not to document the software that you have written. For larger projects you will need to consider other issues such as licenses and distribution. Some advice on creating and maintaining a software project for bioinformatics and computational biology is provided by Stajich et al.⁶⁴ or by Genleman et al.⁶⁵

Perhaps the first decision that must be made is the choice of language(s) and of development environment. There are many different considerations that will affect the choices made. In choosing a language you should consider what tasks need to be carried out, e.g., visualization, machine learning, statistical modeling, data manipulation, string (sequence) matching, and choose a language in which these tasks are easy. You should also consider what languages you are already familiar with, as learning a new language is a time consuming task. Among the many different languages that have found substantial use in this field are Perl, Python, Matlab, S, Java, C, and SQL. There

are also a number of markup languages and protocols such as XML, SOAP, and MAGE-ML that are becoming widely used. We use the term *data technologies* to describe these different markup languages as well as databases in general.

If you intend for others to make use of your software, then no matter how large or small the project there are a number of basic things that you should always do:

- Use version numbers on all software, data etc.
- Document all functions and tasks
- Provide unit testing capabilities.

Version numbers are essential for both software and data. Users will need to be able to determine the cause of changes in their outputs. They will also need to be able to find and obtain newer versions of software and data as they are released. Documentation of both tasks and functions is important so that users can tell both what the software is capable of and whether it is performing as intended. We encourage you to expect software and data to provide these facilities, especially version numbers.

If testing scripts or their equivalent have been provided, then users can determine whether their installation is working correctly. They also allow the developer to add new functionality at a later date and to ensure that existing functionality has not been broken. Finally, and most importantly, testing scripts allow others to extend the software with the confidence that they have not altered any of the core competencies—these users often contribute their ideas back and thereby help to extend the functionality of the software.

In many cases an appropriate analysis of the experimental data will not be possible using preprogrammed analysis tools. In part this is due to the immaturity of the subject area, and in part it is due to the wide range of investigations that are being carried out. In order to adequately analyze data, access to a variety of software tools will be needed, for example, parsing the data, transforming the data, or dropping columns are all common tasks. Becoming familiar with these tools themselves can be a substantial task, and adequate time should be allotted to familiarizing yourself with new software.

One mechanism for delivering software to users is through packages or modules. This is the approach taken in Perl, Matlab, Python, and R. Users of these systems have online resources where they can download a wide variety of software packages. For example, CRAN for R and CPAN for Perl, are software archives that provide the search and distribution facilities. Client-side installation tools are also provided with some languages (e.g., `distutils` for Python and `install.packages` in R). Some languages (such as R, through the `reposTools` package) allow users to set up their own module repositories. Software developers typically upload their modules to an appropriate Web site, and there it is made available for others to download. It is important to check these sites for useful software before starting to write your own because you may find that the tool you need has already been written.

A second mechanism for obtaining computational services is via Web services. Web services are online resources that provide computational support, in some cases fitting models while in others supplying user requested data. To qualify for the

description as a Web service, the resource must support machine generated queries. The explicit structure of the queries is either posted or can itself be accessed interactively via the Web Services Description Language (WSDL). One of the more popular Web services technologies is the Simple Object Access Protocol (SOAP), which is an XML based technology (see the next section for more details on XML). Most high level languages, in particular Perl, Python, and R, have support for XML and SOAP. In the case of R, most of the technological development has been done through the Omegahat project, in particular the R packages **XML**, **SSOAP**, and **RCurl**. There are both client-side and server-side aspects to these technologies; to make use of a Webservice you will need the client-side applications, while to provide a Webservice you will need the server-side software tools.

27.8.1 DATA TECHNOLOGIES

There is an increasing reliance on data technologies in bioinformatics and computational biology. These include the eXtensible Markup Language (XML), which is a technology that allows us to develop vocabularies to describe arbitrarily complex data sets and to share that description with others. Basically this technology provides a mechanism for different software programs to share data. In order to read and correctly interpret data, software requires a description of the data, its format, and to some extent the semantics of the data. XML provides a general solution to many of these problems, and most implementations provide parser support for reading XML files. One should caution that XML is a useful technology for sharing data between software modules when the developer is not in control of both ends of the data pipeline or in cases where it is desirable to make data available to other systems. But when you are in control of all aspects of the program, it can be an extremely inefficient way to move data from one procedure to another.

The size of the data, both experimental and biological, suggests that efficient methods for storing and retrieving data are going to be necessary. This is basically the stock and trade of relational databases. Traditional commercial databases such as Ingres and Oracle have been fairly widely used in support of clinical trials and epidemiological studies. Many bioinformatic projects have been developed on open source databases such as MySQL and PostgreSQL or on lighter weight alternatives such as SQLite, **gdbm** or BerkeleyDB. Each of these alternatives has different strengths and weaknesses, and developers and users should choose one that satisfies their requirements.

Many high level languages, e.g., R, Perl, and Python, provide fairly straightforward tools for accessing many of the popular databases. Accessing databases through ODBC is one solution; in other cases purpose-built interfaces to databases are available. These interfaces have the benefit of often being faster, more efficient, and better able to access special aspects of the database being used than ODBC. On the other hand, they tend to be more difficult to install and require new code bases to be developed should the database change.

A third option is to embed software analysis tools directly into the database. Both PostgreSQL and Oracle support the notion of embedded procedural languages. What this means is that some data manipulation tools can be created and written for

the database itself. This can allow a certain amount of processing to be done within the database itself and can substantially reduce the amount of data that needs to be passed between the database and the processing application.

27.9 DISCUSSION

The analysis of genomic data is an important challenge and a field that is very much in its infancy. Clearly software tools closely aligned with the biological objectives and embodying appropriate statistical methodology will play an essential role. It is unlikely that any one tool will serve all needs or that complex analyses will be possible without some customization of existing software tools. Hence, most substantive research projects will need to add some software engineering capabilities. In essence, one of the effects of the revolution in the ability to study molecular data is the need to expand the analysis team — biologists, statisticians, and software engineers will need to function as a cohesive unit in order to carry out appropriate analyses.

Many of the tools discussed in this chapter will soon be outdated and replaced by new initiatives or better implementations. Practitioners will need to constantly update their software tools and their software skills. Making use of online search engines, new journals aimed at rapid publication, and short course training will help to address this need. As the analyses become more complex, it becomes more difficult to comprehend them and to assess the methods used and the conclusions drawn. The need for reproducible research, in the sense of providing software and data in a format where the analysis can easily be rerun, is an essential ingredient of scientific progress. Without such tools it becomes increasingly difficult to carry out the process of iterative refinement that underlies much of scientific research.

GLOSSARY

The following provides a description and references for some of the software, standards, databases, and projects that are referenced in this chapter. Readers are also referred to the lists provided in some of this chapter's earlier sections for other references.

BerkeleyDB is made by Sleepycat Software, www.sleepycat.com, and is widely-used application-specific data management software.

Bioconductor is an open source project for developing software tools for bioinformatics and computational biology (primarily using R), www.bioconductor.org.

Bioperl is an international association of developers of open source Perl tools for bioinformatics, genomics, and life science research, www.bioperl.org.

Biopython is an international association of developers of freely available Python tools for computational molecular biology, www.biopython.org.

BOOST provides free peer-reviewed portable C++ source libraries, www.boost.org.

caCORE is based at the National Cancer Institute (NCI) and provides a common infrastructure for cancer informatics. The goals of the project are broader than genomic data analysis, although the latter does fall within their domain of interest. See <http://ncicb.nci.nih.gov/core> for more details.

cMAP provides a set of data objects, marked up in XML in such a way that the different pathway components can be extracted and a graph representation of each pathway can be constructed. See <http://cmap.nci.nih.gov/> for more details.

GenePix is a complete stand-alone image analysis software for microarrays, tissue arrays, and cell arrays. It is distributed by Axon Instruments. See <http://www.axon.com/GN>

Graphviz provides a collection of tools for manipulating graph structures and generating graph layouts, www.graphviz.org.

GO The Gene Ontology Consortium (GO) supports and curates three different gene ontologies. They are molecular function, biological process, and cellular component. An ontology is a restricted vocabulary, and in the case of GO that ontology describes genes and gene products. GO provides both the vocabulary (specialized to each of the three ontologies) as well as information about the relationships between the terms. See www.geneontology.org for more details.

GXL The Graph eXchange Language (GXL) is an XML-based representation for describing graphs, i.e., nodes and connecting edges. It is used in a variety of different domains ranging from Bioinformatics to software engineering. See <http://www.gupro.de/GXL/> for more information.

KEGG The Kyoto Encyclopedia of Genes and Genomes⁶⁶ provides a wide variety of data resources. Among their goals is the development of a computationally available representation of the cell and the organism, which enables prediction of cellular processes and behaviors from genomic information. See <http://www.genome.jp/kegg/> for more details.

MATLAB is a fully functional commercial scientific programming language. It is distributed by the MathWorks corporation, <http://www.mathworks.com/>.

MySQL is perhaps the most widely used open source database. It is known for its speed and relatively small size, www.mysql.com.

Octave is an open source implementation of a language that is mostly compatible with MATLAB, www.octave.org.

ODBC Open DataBase Connectivity is a database access standard developed by Microsoft. ODBC enables data between applications and databases. The standard is open and nonproprietary. Information and documentation is available from a variety of sources including the MDAC SDK from Microsoft.

ONCOMINE is a database and integrated data-mining platform for analyzing microarray data that pertain to cancer. Users have access to curated versions of many experiments and can query genes at a variety of resolutions. See www.oncomine.org for more details.

Perl (Practical Extraction and Report Language) is a general high-level programming language that excels at text manipulation. See <http://www.perl.org> and the Comprehensive Perl Archive Network (CPAN) for available add-ons to the system.

PostgreSQL A fully featured open source database. This is a robust database that is widely used, www.postgresql.org.

Python is another high-level scripting language, more structured than Perl with an increasing userbase and collection of contributed extension modules. More information is available from <http://www.python.org>.

R is an open source implementation of the ACM award winning S language and similar to the commercial implementation S-Plus. S is both a general programming language and an extensible interactive environment for data analysis and graphics. See <http://www.r-project.org> for information on the project and CRAN (the Comprehensive R Archive Network) <http://cran.r-project.org> for available software and packages.

SOAP is an acronym for the Simple Object Access Protocol, which is an XML dialect for representing distributed or remote method calls between applications. It has become a very popular protocol for implementing Web services, using HTTP as the communication mechanism and XML as the data representation. See <http://www.w3.org/TR/SOAP/> for more information.

SOURCE is a Web-based database that curates a variety of data (experimental and different biological metadata) and presents an interface that is amenable to the analysis of large microarray datasets. See <http://source.stanford.edu> for more details.

Spot is software for the analysis of microarray images produced by the CSIRO Mathematical and Information Sciences, Image Analysis Group. See <http://experimental.act.cmis.csiro.au/Spot> for more details.

SQLite SQLite is a small implementation of a self-contained, embeddable, SQL database engine, www.sqlite.org.

XML stands for the eXtensible Markup Language, a text-based markup mechanism for representing self-describing data. The W3 organization (<http://www.w3.org>) provides much of the standardization and specification of XML and its dialects. The Cover Pages Web site (<http://xml.coverpages.org>) provides information on using XML in a wide variety of different applications.

REFERENCES

1. Carpenter, AE and Sabatini, DM, Systematic genome-wide screens of gene function, *Nature Reviews Genetics*, 5, 11, 2004.
2. Wang, T-L et al., Digital karyotyping, *PNAS*, 99, 16, 156, 2002.
3. Velculescu, VE, Zhang, L, Vogelstein, B, and Kinzler, K, Serial analysis of gene expression, *Science*, 270, 484, 1995.
4. Ito, T et al., A comprehensive two-hybrid analysis to explore the yeast protein interactome, *PNAS*, 98, 4569, 2001.
5. Gavin, A-C et al., Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature*, 415, 141, 2002.
6. Ho, Y et al., Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry, *Nature*, 415, 180, 2002.
7. Wright, GL, Jr et al., ProteinChip surface enhanced laser desorption/ionization (SELDI) mass spectroscopy: a novel protein biochip technology for detection of prostate cancer biomarkers in complex protein mixtures, *Prostate Cancer and Prostatic Diseases*, 2, 264, 1999.

8. Baggerly, KA, Morris, JS, and Coombes, KR, Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments, computational challenges and solutions, *Bioinformatics*, 20, 777, 2004.
9. R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004.
10. Cope, L et al., A benchmark for Affymetrix GeneChip expression measures, *Bioinformatics*, 20, 323, 2004.
11. Edgar, R, Domrachev, M, and Lash, A, Gene expression omnibus: NCBI gene expression and hybridization array data repository, *Nucleic Acids Research*, 30, 207, 2002.
12. Brazma, A et al., ArrayExpress: a public repository for microarray gene expression data at the EBI, *Nucleic Acids Research*, 31, 68, 2003.
13. Cheung KH et al., YMD: microarray database for large-scale gene expression analysis, *Proceedings of the American Medical Informatics Association*, 140, 2002.
14. Parmigiani, G, Garrett, E, Irizarry, R, and Zeger, S, eds., *The Analysis of Gene Expression Data: Methods and Software*, Springer, New York, 2003.
15. Speed, TP, ed., *Statistical Analysis of Gene Expression Microarray Data*, Chapman and Hall, 2003.
16. Zhang, W, Shmulevich, I, and Astola, J, *Microarray Quality Control*, John Wiley and Sons, New York 2004.
17. Simon, RM et al., *Design and Analysis of DNA Microarray Investigations*, Springer, New York, 2004.
18. Jain, AN et al., Fully automatic quantification of microarray image data, *Genome Research*, 12, 325, 2002.
19. Yang, YH, Buckley, MJ, Dudoit, S, and Speed, TP, Comparison of methods for image analysis on CDNA microarray data, *Journal of Computational and Graphical Statistics*, 11, 108, 2002.
20. Tadesse, MG, Ibrahim, JG, and Gentleman, R, Bayesian error-in-variables survival model for the analysis of GeneChip arrays, *Biometrics*, to appear.
21. Irizarry, R, Gautier, L, and Cope, L, An R package for analyses of Affymetrix oligonucleotide arrays, in Parmigiani, G, Garrett, ES, Irizarry, RA, and Zeger, SL eds., *The Analysis of Gene Expression Data: Methods and Software*, Springer, New York, 2003.
22. Wu, Z et al., A model based background adjustment for oligonucleotide expression arrays, *JASA*, to appear, 2004.
23. Dudoit, S and Yang, YH, Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data, in Parmigiani, G, Garrett, ES, Irizarry, RA, and Zeger, SL, eds., *The Analysis of Gene Expression Data: Methods and Software*, Springer, New York, 2003.
24. Huber, W et al., Variance stabilization applied to microarray data calibration and to the quantification of differential expression, *Bioinformatics*, 18 Suppl. 1, S96, 2002.
25. von Heydebreck, A, Huber, W, and Gentleman, R, Differential expression with the bioconductor project, in *Encyclopedia of Genetics, Genomics, Proteomics, and Bioinformatics*, John Wiley and Sons, New York 2004.
26. Smyth, G, Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, *Statistical Applications in Genetics and Molecular Biology*, 3, article 3, 2004.
27. Mootha, VK et al., PGC-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes, *Nature Genetics*, 34, 267, 2003.
28. Kendzierski, C, Newton, M, Lan, H, and Gould, M, On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles, *Statistics in Medicine*, 22, 3899, 2003.

29. Baldi, P and Long, A, A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, 17, 509, June 2001.
30. Tusher, V, Tibshirani, R, and Chu, G, Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Science*, 98, 5116, 2001.
31. Lönnstedt, I and Speed, TP. Replicated microarray data, *Statistica Sinica*, 12, 31, 2002.
32. Li, C and Wong, WH, Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection, *Proceedings of the National Academy of Science*, 98, 31, 2001.
33. Newton, M et al., On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data, *Journal of Computational Biology*, 8, 37, 2001.
34. Wu, H, Kerr, MK, Cui, X, and Churchill, GA, MAANOVA: a software package for the analysis of spotted cDNA microarray experiments, 313, in Parmigiani, G, Garrett, E, Irizarry, R, and Zeger, S, eds., *The Analysis of Gene Expression Data: Methods and Software*, Springer, New York, 2003.
35. Yasui, Y et al., A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection, *Bioinformatics*, 20, 777, 2004.
36. Petricoin, EF et al., Use of proteomic patterns in serum to identify ovarian cancer, *Lancet*, 359, 572, 2002.
37. Snijders, AM et al., Assembly of microarrays for genome-wide measurement of DNA copy number, *Nature Genetics*, 29, 263, 2001.
38. Rhodes, D R et al., ONCOMINE: a cancer microarray database and integrated data-mining platform, *Neoplasia*, 6, 1, 2004.
39. Diehn, M et al., SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data, *Nucleic Acids Research*, 31, 219, 2003.
40. Saal, LH et al., Bioarray software environment (BASE): a platform for comprehensive management and analysis of microarray data, *Genome Biology*, 3, software0003.1, 2002.
41. Covitz, PA et al., caCORE: a common infrastructure for cancer informatics, *Bioinformatics*, 19, 2404, 2003.
42. Parmigiani, G, Garrett-Mayer, E, Anbazhagan, R, and Gabrielson, E, A cross-study comparison of gene expression studies for the molecular classification of lung cancer, *Clinical Cancer Research*, 10, 2922, 2004.
43. Choi, JK, Yu, U, Kim, S, and Yoo, OJ, Combining multiple microarray studies and modeling interstudy variation, *Bioinformatics*, 19, Suppl. 1, i84, 2003.
44. Hastie, T, Tibshirani, R, and Friedman, JH, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Verlag, Berlin 2001.
45. Venables, W and Ripley, B, *Modern Applied Statistics with S*, 4th ed., Springer, New York, 2002.
46. Duda, RO, Hart, PE, and Stork, DG, *Pattern Classification*, John Wiley and Sons, New York 2001.
47. Ripley, B, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, 1996.
48. Törönen, P, Kolehainen, M, Wong, G, and Castren, E, Analysis of gene expression data using self-organizing maps, *FEBS Letters*, 451, 142, 1999.
49. Kaufman, L and Rousseeuw, P, *Finding Groups in Data*, John Wiley and Sons, New York 1990.

50. Eisen, MB, Spellman, PT, Brown, PO, and Botstein, D, Cluster analysis and display of genome-wide expression patterns, *PNAS*, 95, 14,863, 1998.
51. Ding, B and Gentleman, R. Testing gene associations using co-citation. Technical report, The Bioconductor Project, 2004.
52. Breiman, L, Random forests, *Machine Learning*, 45, 5, 2001.
53. Tanay, A, Sharan, R, and Shamir, R, Discovering statistically significant biclusters in gene expression data, in *Proceedings of the Tenth International Conference on Intelligent Systems for Molecular Biology*, 136, Bioinformatics, New York, 2002.
54. Balasubramanian, R, LaFramboise, T, Scholtens, D, and Gentleman, R, A graph theoretic approach to testing associations between disparate sources of functional genomics data, *Bioinformatics*, 20(18), 3353–3362, 2004.
55. Ge, H, Liu, Z, Church, G, and Vidal, M, Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*, *Nature Genetics*, 29, 482, 2001.
56. Melhorn, K and Näher, S, *LEDA*, Cambridge University Press, Cambridge 1999.
57. Siek, JG, Lee, LQ, and Lumsdaine, A, *The Boost Graph Library*, Addison Wesley, New York, 2002.
58. Gansner, E, Koutsofios, E, North, SC, and Vo, K-P, A technique for drawing directed graphs. Technical report, 2003.
59. David, A and Mathieu, B, *Tulip: a system dedicated, to the visualization of huge graphs*, 2004. <http://www.tulip-software.org>.
60. Rossini, A, Literate statistical analysis, in Hornik, K and Leisch, F, eds., *Proceedings of the 2nd International Workshop on Distributed Statistical Computing*, March 15–17, 2002., <http://www.ci.tuwien.ac.at/Conferences/DSC-2001/Proceedings>, 2001.
61. Gentleman, R and Temple Lang, D, Statistical analyses and reproducible research, to appear.
62. Robert Gentleman (2005) “Reproducible Research: A Bioinformatics Case Study”, *Statistical Applications in Genetics and Molecular Biology*, 4(1), Article 2. <http://www.bepress.com/sagmb/vol4/issi/arts>
63. Leisch, F, Sweave: Dynamic generation of statistical reports using literate data analysis, 575, in Härdle, W and Rönz, B, eds., *Compstat 2002: Proceedings in Computational Statistics*, Physika Verlag, Heidelberg, Germany, 2002.
64. Stajich, J et al., The Bioperl toolkit: Perl modules for the life sciences, *Genome Research*, 12, 1611, 2002.
65. Gentleman, RC et al., Bioconductor: open software development for computational biology and bioinformatics, *Genome Biology*, 5:R80, doi:10.1186/gb, 2004.
66. Kanehisa, M and Goto, S, KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Research*, 28, 27, 2000.

Part VI

Interpreting Clinical Trials

28 Bayesian Sensitivity Analyses of Confounded Treatment Effects

Peter F. Thall and Xuemei Wang

CONTENTS

28.1	Introduction	523
28.2	Survival Analysis with Treatment-Trial Confounding	526
28.2.1	Comparing Leukemia Treatments	526
28.2.2	Probability Models	527
28.2.3	Sensitivity Analyses of the Leukemia Data	527
28.3	Analyzing Confounded Count Data	533
28.3.1	Data and Problem Definition	533
28.3.2	Probability Model	535
28.3.3	Sensitivity Analyses of the Transplant Data	536
28.4	Discussion	539
	Acknowledgment	539
	References	539

28.1 INTRODUCTION

The primary scientific goal of a randomized clinical trial of two treatments, A and B, is to compare their effects on the most important therapeutic outcome in the medical setting under study. Generally, this comparison may be formulated in terms of two real-valued parameters, θ_A and θ_B , which most often are based on probabilities or hazards under treatments A and B, respectively, possibly transformed or covariate-adjusted. Scientists routinely base such comparisons on the A-versus-B treatment effect, $\delta_\theta = \theta_A - \theta_B$, implicitly assuming that a typical statistical estimator computed from their data actually estimates δ_θ . While it is well established that randomization will, on average, eliminate potential sources of bias,^{1,2} when patients are not randomized between A and B, standard statistical estimators may become scientifically invalid and substantively misleading.

Even with randomization, in practice there are many difficulties. Patient outcome in any medical setting arises from the combined effects of treatment, patient characteristics, and other “latent” variables that are unobserved. It is well known that

failure to account for patient covariates that affect outcome may produce a biased estimate of δ_θ .³ However, after accounting for known covariates through stratification, dynamic randomization,⁴ or regression analysis, between-patient variability still may be substantial. Because latent variables are not observed, their combined effect may appear to be attributable to treatment, thus masking an actual treatment effect. Potential sources of such systematic bias include supportive care, physician practices, or institutional effects. The randomization itself may fail in numerous ways. For example, patients who enter an unblinded trial in the hope of being treated with A may drop out if randomized to B. An adverse event more likely with A than with B may cause higher rates of dropout or early treatment termination in arm A. Physicians who favor one treatment over the other may, consciously or unconsciously, selectively withhold patients from the trial. In small to moderate sized studies, covariate imbalance may be substantial simply due to the play of chance. Still, despite the many problems imposed by practical reality, randomization remains the best scientific device for obtaining a fair treatment comparison.

The statistical problem of evaluating δ_θ becomes much more difficult when the data arise from separate trials of A and B. If A is evaluated in trial 1 and B is evaluated in trial 2, then the data from trial 1 provide an estimator not of θ_A but rather of the confounded effect $\gamma_{A,1}$ of treatment A and trial 1. Similarly, the data from trial 2 provide an estimator of the confounded effect $\gamma_{B,2}$ of treatment B and trial 2. A typical estimator of δ_θ based on such data actually estimates the difference between these confounded effects $\delta = \gamma_{A,1} - \gamma_{B,2}$ rather than the treatment effect δ_θ . Of course, the distinction between δ and δ_θ is the motivation for conducting a randomized trial. There are numerous examples of between-trial effects in the statistical and medical literature. See, for example, Estey and Thall.⁵

Unfortunately, it is common practice throughout the medical literature to base statistical comparisons on data from nonrandomized experiments while ignoring study effects. A typical report of a single-arm trial of a treatment A may compare it to another treatment B based on data from one or more previous studies of B. This scientific problem is prominent in the reported results of many phase II clinical trials, which by convention are most often conducted as single-arm studies. Evidently, many scientists believe that the use of regression methods or subset analyses to account for patient covariates when making such comparisons provides a valid estimate of δ_θ . The underlying assumptions, usually not stated explicitly, are that known covariates account for any between-trial effects,^{6,7} or that between-trial effects either are negligible relative to δ_θ or simply do not exist. This explains, in part, why many randomized phase III trials yield negative results despite the fact that one or more preceding single-arm phase II trials indicated that the experimental treatment under study was promising compared to standard therapy.

This chapter has two purposes. The first is to show by example that substantial between-trial effects may persist after accounting for known covariates. We will illustrate by example the fact that between-trial effects may be substantial even when the trials are conducted in the same institution and one accounts for covariate effects. The second purpose is to illustrate some simple Bayesian methods for assessing what the possible distribution of δ_θ may be in such settings. Because from the Bayesian viewpoint all parameters are random quantities, we will focus on the problem of

hypothesizing what the posterior distribution of δ_θ may be when it cannot actually be computed from the available data. The basic idea underlying our approach is as follows. Let D_A and D_B denote the data from the separate trials of A and B, and denote by $f(\delta | D_A, D_B)$ the posterior distribution of the confounded effect $\delta = \gamma_{A,1} - \gamma_{B,2}$. Denoting the trial effects by λ_1 and λ_2 , we assume that $\gamma_{A,1} = \theta_A + \lambda_1$, the sum of effects due to treatment A and trial 1 and, similarly, that $\gamma_{B,2} = \theta_B + \lambda_2$. This implies that

$$\delta = (\theta_A - \theta_B) + (\lambda_1 - \lambda_2) = \delta_\theta + \delta_\lambda.$$

That is, δ equals the sum of the A-versus-B treatment effect of interest $\delta_\theta = \theta_A - \theta_B$ and the confounding between-trial effect $\delta_\lambda = \lambda_1 - \lambda_2$. The problem is that based on the available data the posterior of δ can be computed, but the posteriors of δ_θ and δ_λ cannot be computed. That is, δ_θ and δ_λ are confounded. The problem of confounding still exists if one adopts a frequentist rather than a Bayesian viewpoint. Under a frequentist formulation, in which the parameters are considered to be fixed but unknown quantities, based on the available data the distribution of δ is identifiable, but any distribution including δ_θ and δ_λ as distinct parameters is not identifiable. Consequently, from a frequentist viewpoint it is not possible to estimate δ_θ without including additional data to estimate δ_λ and also making some rather strong additional assumptions. We will revisit this issue in the context of our first illustration. The simple decomposition $\delta = \delta_\theta + \delta_\lambda$ reveals the assumption in a typical analysis that ignores the fact that patients were not randomized between A and B because assuming that $E(\delta) = E(\delta_\theta)$ is equivalent to assuming that $E(\lambda_1) = E(\lambda_2)$; that is, that the mean trial effects are identical. This also underlies the approach of Begg and Pilote,⁸ and Li and Begg,⁹ who consider meta-analysis of combined single-arm and randomized study data. In particular, they assume that the between-trial effects for single-arm trials all have a common mean. We do not make such an assumption here.

Our first type of sensitivity analysis is carried out by varying a hypothetical distribution for δ_λ over a reasonable set of possibilities and assessing the resulting distributions of $\delta_\theta = \delta - \delta_\lambda$. When the only available data are D_A and D_B from two separate trials of A and B, our approach is to compute the posterior $f(\delta | D_A, D_B)$, hypothesize a distribution $f^{(h)}(\delta_\lambda | D_A, D_B)$ for the between-trial effect, and use these two distributions to compute a *hypothetical posterior* $f^{(h)}(\delta_\theta | D_A, D_B)$ for the treatment effect. We consider it essential to refer to this as a hypothetical posterior because the actual posterior $f(\delta_\theta | D_A, D_B)$ of δ_θ cannot be computed unless one makes the assumption that the distribution of the between-trial effect δ_λ is known. If one assumes that δ_θ and δ_λ are independent, then $\text{var}(\delta) = \text{var}(\delta_\theta) + \text{var}(\delta_\lambda)$, and the problem of specifying $f^{(h)}(\delta_\lambda | D_A, D_B)$ may be simplified by assuming that $\text{var}(\delta_\lambda) = 1/2 \text{var}(\delta)$. Given $\text{var}(\delta_\lambda)$, assuming normality, one may then obtain a range of distributions for δ_λ by varying its hypothetical mean, with the sensitivity analysis consisting of computing $\text{pr}^{(h)}(\delta_\theta > 0 | D_A, D_B)$ as a function of the hypothetical mean of δ_λ . In practice, any reasonable value of $\text{var}(\delta_\lambda)$ near the magnitude of $\text{var}(\delta)$ gives the same substantive conclusions. If desired, one may take a more refined version of this approach by varying both the mean and the variance of δ_λ .

When historical data D from one or more other clinical settings similar to the trials of A and B are available, one may vary hypothetical $f^{(h)}(\delta_\lambda | D_A, D_B, D)$ over a set of historical between-trial effects. While this relies on the assumption that δ_λ and the historical between-trial effects are stochastically similar, we do not assume that the posteriors of either δ_λ or δ_θ may be computed from historical data. As in the simpler case described above where the only available data are D_A and D_B , we refer to $f^{(h)}(\delta_\theta | D_A, D_B, D)$ as the *hypothetical posterior* of δ_θ .

More specifically, to perform this type of sensitivity analysis in our first application we include data from four additional trials. In that analysis, the two treatments we wish to compare, A and B, were studied in trials 1 and 2, respectively, while trials 3 and 4 both studied a third treatment, C, and trials 5 and 6 studied a fourth treatment, D. Using trial 1 as the baseline with δ_j the comparative effect of trial j versus trial 1 for $j = 2, \dots, 6$ and $\delta_1 = 0$, we assume that $\delta_2 = \delta_\theta + \delta_{\lambda,2}$, $\delta_3 = \delta_C + \delta_{\lambda,3}$, $\delta_4 = \delta_C + \delta_{\lambda,4}$, $\delta_5 = \delta_D + \delta_{\lambda,5}$ and $\delta_6 = \delta_D + \delta_{\lambda,6}$. Thus, δ_θ is the B-versus-A treatment effect of interest, δ_C is the C-versus-A effect, and δ_D is the D-versus-A effect. Given the additional data from trials 3–6, we use the actual posteriors of the between-trial effects, $\delta_{\lambda,3} - \delta_{\lambda,4} = \delta_3 - \delta_4$ and $\delta_{\lambda,5} - \delta_{\lambda,6} = \delta_5 - \delta_6$, as hypothetical posteriors for $\delta_{\lambda,2} - \delta_{\lambda,1} = \delta_{\lambda,2}$.

In this setting, one may make the stronger assumption that the between-trial effects $\delta_2 - \delta_1$, $\delta_3 - \delta_4$, and $\delta_5 - \delta_6$ are identically distributed, use the posterior of $\delta_3 - \delta_4$ or $\delta_5 - \delta_6$, or possibly some weighted average, as the prior for $\delta_{\lambda,2}$ and integrate $\delta_{\lambda,2}$ out of $p(\delta_{\lambda,2}, \delta | \text{data})$ under the assumption that $\delta_2 = \delta_\theta + \delta_{\lambda,2}$ to obtain a posterior $p(\delta_\theta | \text{data})$ for the treatment effect of interest. If one is willing to make such stronger assumptions, then in general given any historical data on between-trial effects one would never need to randomize because $p(\delta_\theta | \text{data})$ could always be obtained from the historical data or, more generally, by simply assuming a prior on δ_λ and computing the resulting posterior of δ_θ . A frequentist analog of this approach would use historical data to estimate δ_λ and then subtract this estimate from that of δ to obtain an estimate of δ_θ . The point is that any probability distribution that treats δ_θ and δ_λ as distinct parameters is not identifiable due to treatment-trial confounding. Moreover, while data on between-trial effects in other studies may be similar to putative, nonexistent data on the effects of trials A and B, the assumption that historical data may be substituted for such nonexistent data is quite strong. We prefer the more conservative interpretation that the resulting posterior of δ_θ is hypothetical. For either type of sensitivity analysis, hypothesizing a number of different distributions for the unknown posterior of δ_λ and computing the resulting posteriors of δ_θ is analogous to the more conventional Bayesian procedure of varying the prior and assessing the effect on posterior quantities.

28.2 SURVIVAL ANALYSIS WITH TREATMENT-TRIAL CONFOUNDING

28.2.1 COMPARING LEUKEMIA TREATMENTS

Our first illustration is motivated by the desire to compare the efficacy of two combination chemotherapies for patients 65 years or older with newly diagnosed acute

myelogenous leukemia (AML) or myelodysplastic syndromes (MDS). Survival time and baseline covariate data were available for patients from two separate single-arm trials, both conducted at M.D. Anderson Cancer Center (MDACC). The first was a trial conducted in 1991–92 of the well-established combination idarubicin + cytosine arabinoside, “ara-C” (IA), and the second a subsequent trial of the newer combination gemtuzumab ozogamicin (GO). Patients in the GO trial were randomized between GO and GO plus interleukin 11 (IL-11). Additional details are given by Estey et al.¹⁰ Because the IL-11 effect was negligible, we collapse the arms of the GO trial and focus on evaluating the GO-versus-IA effect, δ_{GO} . Because the AML/MDS patients were not randomized between GO and IA, any conventional statistical estimator of δ_{GO} is confounded by the between-trial effect.

28.2.2 PROBABILITY MODELS

The patient covariates included in our survival analyses are Zubrod performance status (PS), dichotomized as “good” = [PS ≤ 2] versus “poor” = [PS ≥ 3]; whether the patient was treated in a laminar airflow room (LAR); and cytogenetic karyotype, classified into three categories: normal (the baseline group), having the very unfavorable –5/–7 abnormality or having an abnormality other than –5/–7. We denote the linear combination of these covariates by $\beta\mathbf{Z} = \beta_0 + \beta_1Z_1 + \dots + \beta_4Z_4$, so that β_0 is the baseline hazard parameter corresponding to $Z_1 = Z_2 = Z_3 = Z_4 = 0$. Let $S(t|\mathbf{Z}) = \text{pr}(T > t | \mathbf{Z}, \boldsymbol{\theta})$ be the survivor function, where T denotes survival time and $\boldsymbol{\theta}$ is the model parameter vector.

Initially, we considered three possible models for survival time. Let $\delta\boldsymbol{\tau}$ denote a linear combination of one or more confounded treatment-trial effects using trial 1 as the baseline. The three models were the Weibull, for which $\log[-\log\{S(t|\mathbf{Z})\}] = \beta\mathbf{Z} + \delta\boldsymbol{\tau} + \phi\log(t)$, the log logistic, for which $-\log[S(t|\mathbf{Z})/\{1-S(t|\mathbf{Z})\}] = \beta\mathbf{Z} + \delta\boldsymbol{\tau} + \phi\log(t)$, and the lognormal with mean $\beta\mathbf{Z} + \delta\boldsymbol{\tau}$ and constant variance. For the Weibull and log logistic, ϕ is a shape parameter that determines how the hazard of death may vary over time. For the AML/MDS data from all six trials, these models have respective maximized log likelihoods –137.0, –139.4, and –141.7, indicating that the Weibull gives a slightly better fit. Additionally, a plot of $\log[-\log\{S_{KM}(t)\}]$ on $\log(t)$ (not shown), where $S_{KM}(t)$ is the Kaplan–Meier estimator,¹¹ is approximately linear, indicating that the Weibull assumption is reasonable. We thus will use this model for our sensitivity analyses.

28.2.3 SENSITIVITY ANALYSES OF THE LEUKEMIA DATA

We will perform two different types of sensitivity analyses. The first is based on only the data from the GO and IA trials, and we obtain hypothetical posterior treatment effects by varying the mean of a hypothetical distribution of the confounding between-trial effect. That is, first identifying the GO-versus-IA treatment effect $\delta_\theta = \delta_{GO}$ in the notation of Section 28.1, we vary the mean of the hypothetical between-trial effect δ_λ and evaluate the resulting hypothetical posterior of the GO-versus-IA treatment effect $\delta_{GO} = \delta - \delta_\lambda$. The second analysis includes additional data from four trials of other treatments for AML/MDS conducted at MDACC. Since our focus is the

GO-versus-IA treatment effect, the data from these four additional trials are included only to provide a basis for constructing hypothetical trial effect distributions.

For our first sensitivity analysis, based on the data from the GO and IA trials, we fit the Weibull regression model characterized by $\log[-\log\{S(t|\mathbf{Z})\}] = \beta\mathbf{Z} + \delta\tau + \phi\log(t)$, where τ is the indicator of the GO trial and thus δ is the confounded effect of GO in trial 2 versus IA in trial 1. For priors in both this analysis and the fit of the extended Weibull model described below, we assumed that ϕ followed a gamma distribution with mean = 1 and variance = 10^4 and that all other parameters were normal with mean 0 and variance 10^5 . Posteriors were computed in WinBugs version 1.4. After an initial burn-in of 1,000 Markov chain Monte Carlo (MCMC) iterations, convergence was checked using the diagnostic tool of Brooks and Gelman.¹² A thinning interval of 10 was used, with each posterior based on 1,000 MCMC samples from 10,000 iterations.

The fitted model is summarized in Table 28.1, which shows that the effects of PS and LAR were equivocal because the parameter of each has 95% posterior credible interval, running from the 2.5th to the 97.5th percentiles of the posterior, containing 0. The type of cytogenetic abnormality was highly predictive of survival, with the relative risk (RR) of death in the $-5/-7$ group and the “other cytogenetic abnormalities” group on average about 4.5 times and 3.4 times that of the group with normal cytogenetics, respectively. The posterior mean of δ was 0.84, corresponding to $RR = 2.3$, which indicates that patients in the GO trial had substantially worse survival than those in the IA trial. To perform the sensitivity analysis, we make the key assumptions that $\delta = \delta_{GO} + \delta_{\lambda,2}$, where δ_{GO} is the true GO-versus-IA treatment effect and $\delta_{\lambda,2}$ is the trial 2-versus-trial 1 effect, that $\delta_{\lambda,2}$ is normally distributed, and that $\text{var}(\delta) = \text{var}(\delta_{GO}) + \text{var}(\delta_{\lambda,2})$ with $\text{var}(\delta_{\lambda,2}) = \text{var}(\delta_{GO}) = 1/2 \text{var}(\delta) = 0.065$. If $E(\delta_{\lambda,2}) = 0$ corresponding to on average no trial effect, then $E(\delta_{GO}) = E(\delta) = 0.84$, so that on average all of the observed effect is due to treatment GO-versus-IA. If $E(\delta_{\lambda,2}) = 0.84$ so that on average all of the mean effect was due a between-trial difference, then $E(\delta_{GO}) = 0$, and on average none of the observed effect is due to

TABLE 28.1

Fitted Bayesian Weibull Regression Model for Survival Time in the IA and GO Trials. Aside from ϕ , a positive parameter value is associated with shorter survival

Variable	Posterior Distribution	
	Mean _{SD}	95% Credible Interval
Intercept	-1.14 _{0.39}	(-1.93, -0.43)
Performance status = 3 or 4	0.24 _{0.39}	(-0.55, 0.96)
Treatment in laminar airflow room	-0.43 _{0.34}	(-1.09, 0.22)
$-5/-7$ cytogenetic abnormality	1.52 _{0.43}	(0.64, 2.43)
Other cytogenetic abnormalities	1.22 _{0.40}	(0.46, 2.06)
GO in trial 2 versus IA in trial 1	0.84 _{0.36}	(0.14, 1.53)
Shape parameter (ϕ)	0.83 _{0.09}	(0.66, 1.02)

treatment. Our sensitivity analysis consists of varying $E(\delta_{\lambda,2})$ over the domain from 0 to 0.84 and computing the resulting values of $\text{pr}^{(h)}(\delta_{\text{GO}} > 0 \mid \text{data}) = \text{pr}^{(h)}(\text{survival is worse with GO} \mid \text{data})$. A plot of these hypothetical values is given in Figure 28.1, with the values of $E(\delta_{\lambda,2})$ replaced by their percentages of 0.84. Figure 28.1 indicates that, based on these data and under the above assumptions, even if up to 60% of the confounded effect is due to between-trial differences, then it is still nearly certain that survival is worse with GO than with IA.

For our second sensitivity analysis, we incorporated additional data from four other single-arm trials in AML/MDS, also conducted at MDACC over the same time period, in order to obtain hypothetical between-trial effects. Two of the four trials were of fludarabine + IA + G-CSF (FAIG), and the other two were trials of FAIG + all-trans retinoic acid (FAIGA). Figure 28.2 gives the Kaplan–Meier estimates of the six survival curves. These unadjusted survival estimates indicate that patients in the GO trial and one FAIG trial had the worst survival, although the GO trial had by far the shortest follow-up. The curves also illustrate the well-known fact that in this older AML/MDS patient group survival is poor for any type of chemotherapy. It also must be kept in mind that comparisons among these six curves do not correspond to the treatments IA, GO, FAIG, and FAIGA per se, but to the six confounded treatment-trial effects because patients were not randomized among the treatments.

Table 28.2 gives the distributions of the covariates within each of the six trials. Each covariate shows nearly the same distribution across trials. The exceptions are that the IA trial has a lower percentage of patients treated in the LAR; the two FAIG

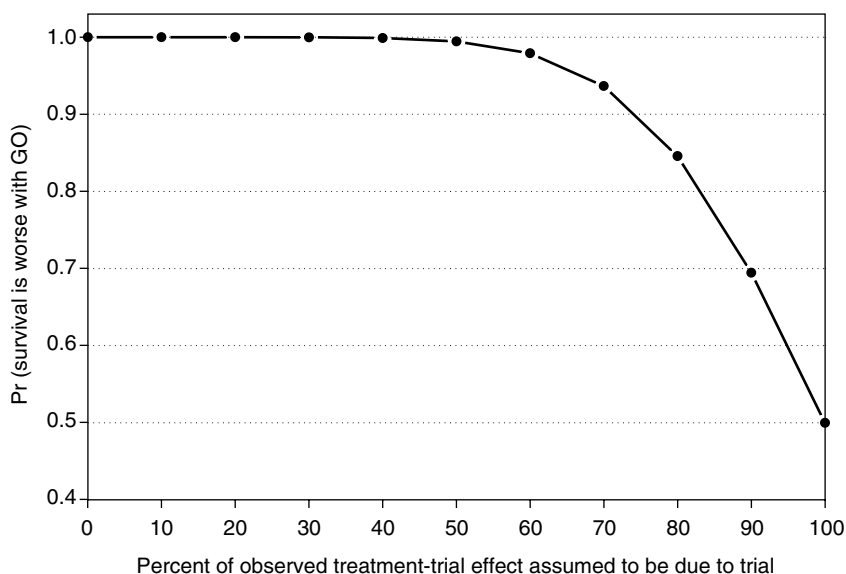


FIGURE 28.1 Hypothetical probability that survival with GO is worse than with IA as a function of the assumed mean of the hypothetical between-trial effect, $\delta_{\lambda,2}$. On the horizontal axis, 0% corresponds to $E(\delta_{\lambda,2}) = 0$ and 100% to $E(\delta_{\lambda,2}) = 0.84$, the posterior mean of the confounded treatment-trial effect.

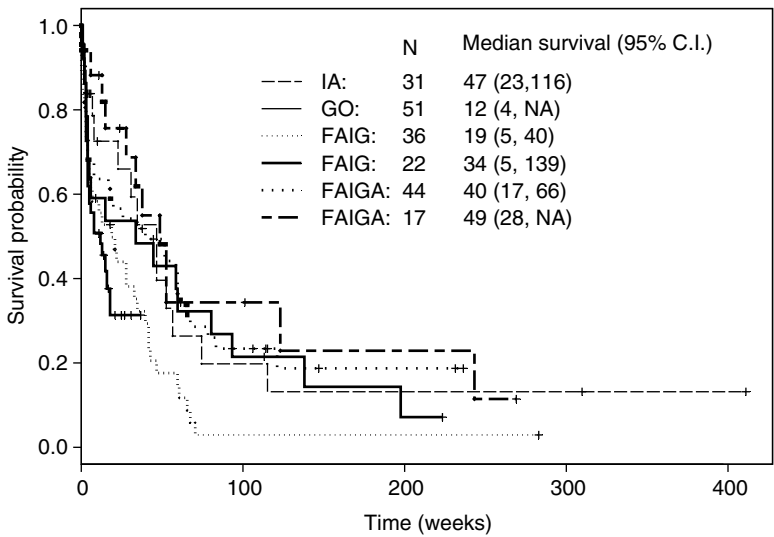


FIGURE 28.2 Kaplan–Meier plots and summary statistics of the survival distributions in the six AML/MDS chemotherapy trials.

TABLE 28.2
Distribution of Covariates within Each of the Six Trials. For each covariate in each trial, the count is followed in parentheses by the corresponding percentage

Trial	# Patients	PS = 3, 4	LAR	Cyto = -5/-7	Cyto = Other
IA	31	4 (12.9)	13 (41.9)	8 (25.8)	11 (35.5)
GO	51	8 (15.7)	38 (74.5)	15 (29.4)	16 (31.4)
FAIG	36	5 (13.9)	30 (83.3)	18 (50.0)	17 (47.2)
FAIG	22	4 (18.2)	18 (81.8)	5 (22.7)	10 (45.5)
FAIGA	44	6 (13.6)	35 (79.5)	13 (29.5)	14 (31.8)
FAIGA	17	1 (5.9)	12 (70.6)	7 (41.2)	5 (29.4)

trials have slightly higher percentages of patients with “other” cytogenetics as well as the highest and lowest percentages of -5/-7; and only 1 of the 17 patients in trial 6 has PS≥3.

In the component $\delta\tau = \delta_2\tau_2 + \dots + \delta_6\tau_6$ of the linear term of the Weibull model that accounts for trial effects, for $j = 2, \dots, 6$, the j th treatment-trial combination compared to IA in trial 1 is δ_j and the trial indicator is τ_j . We fit the extended model that includes $\delta\tau$ to the data from all six trials using priors and MCMC methods similar to those used to obtain the earlier fit of the simpler model in Table 28.1. The fitted model summarized in Table 28.3 leads to the same substantive conclusions regarding covariate effects as the fit based on only trials 1 and 2, although the standard deviations are smaller reflecting the larger overall sample size.

TABLE 28.3
Fitted Bayesian Weibull Regression Model for Survival Time in the Six Trials: IA in Trial 1, GO in Trial 2, FAIG in Trials 3 and 4, and FAIG + ATRA in Trials 5 and 6. Aside from ϕ , a positive parameter value is associated with shorter survival

Variable	Posterior Distribution	
	Mean _{sd}	95% Credible Interval
Intercept	-0.62 _{0.29}	(-1.21, -0.09)
Performance status = 3 or 4	0.67 _{0.24}	(0.20, 1.14)
Treatment in laminar airflow room	-1.04 _{0.21}	(-1.45, -0.61)
-5/-7 cytogenetic abnormality	1.23 _{0.24}	(0.77, 1.71)
Other cytogenetic abnormalities	0.63 _{0.24}	(0.18, 1.09)
Shape parameter (ϕ)	0.71 _{0.05}	(0.63, 0.81)
Treatment-trial effects, versus IA in trial 1		
GO in trial 2 (δ_2)	0.84 _{0.33}	(0.20, 1.48)
FAIG in trial 3 (δ_3)	0.74 _{0.35}	(0.10, 1.45)
FAIG in trial 4 (δ_4)	0.47 _{0.38}	(-0.27, 1.23)
FAIG + ATRA in trial 5 (δ_5)	0.39 _{0.34}	(-0.25, 1.08)
FAIG + ATRA in trial 6 (δ_6)	-0.21 _{0.40}	(-0.96, 0.61)

Our second sensitivity analysis is based on this extended data set, and it relies on the fitted model in Table 28.3. We will assume that each of the treatment-trial effects for trials 2, 3, 4, 5, and 6 compared to trial 1 may be decomposed into the sum of a treatment effect and a trial effect. Formally, we assume that $\delta_2 = \delta_{GO} + \delta_{\lambda,2}$, $\delta_3 = \delta_{FAIG} + \delta_{\lambda,3}$, $\delta_4 = \delta_{FAIG} + \delta_{\lambda,4}$, $\delta_5 = \delta_{FAIGA} + \delta_{\lambda,5}$ and $\delta_6 = \delta_{FAIGA} + \delta_{\lambda,6}$. This allows us to exploit the fact that FAIG and FAIGA were each studied in two separate trials. That is, $\delta_3 - \delta_4 = \delta_{\lambda,3} - \delta_{\lambda,4}$ and $\delta_5 - \delta_6 = \delta_{\lambda,5} - \delta_{\lambda,6}$ are actual covariate-adjusted between-trial effects. Because the signs of these two differences are artifacts of the way the trials were indexed, $\delta_4 - \delta_3$ and $\delta_6 - \delta_5$ also are between-trial effects. The posteriors of these four between-trial effects are plotted in Figure 28.3.

Based on the fit of the extended model, the fact that $\text{pr}(\delta_2 > 0 \mid \text{data}) > 0.99$ still indicates that patients given GO in trial 2 had substantially worse survival than those given IA in trial 1. Under the assumption that $\delta_2 = \delta_{GO} + \delta_{\lambda,2}$, we again address the question of how much of δ_2 may have been due to the GO-versus-IA treatment effect δ_{GO} , but we use the actual between-trial effects from the FAIG and FAIG + ATRA trials as the hypothetical trial 2-versus-trial 1 effect $\delta_{\lambda,2}$. For the two FAIG trials, the posterior probability of a positive trial effect equals either 0.18 or 0.82 depending upon whether one computes $\text{pr}(\delta_{\lambda,4} - \delta_{\lambda,3} > 0 \mid \text{data})$ or $\text{pr}(\delta_{\lambda,3} - \delta_{\lambda,4} > 0 \mid \text{data})$. For the two FAIGA trials, these probabilities are either 0.04 or 0.96. This illustrates the fact that, even after accounting for the effects of known covariates, remaining between-trial effects may be substantial. When one ignores a between-trial effect such as those in Figure 28.3, as is routinely done in the medical literature, this effect is added to any actual treatment effect that may exist, and the sum is incorrectly considered to be the treatment effect. These data provide an explicit illustration of

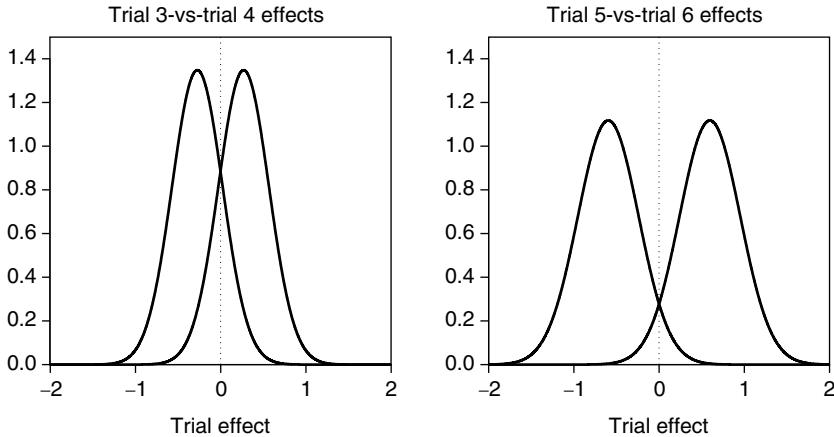


FIGURE 28.3 Posterior between-trial effect distributions for the two FAIG trials (3 and 4) and the two FAIGA trials (5 and 6).

between-trial effects because, if one were not told that trials 5 and 6 studied the same treatment but instead assumed that two different treatments were studied, a routine data analysis would lead to the conclusion that the treatment in trial 6 was greatly superior to that in trial 5.

For the second sensitivity analysis, first note that $\delta_{\lambda,2}$ is the trial 2-versus-trial 1 effect. We successively assume that the hypothetical distribution of $\delta_{\lambda,2}$ is that of the trial 3-versus-trial 4 effect, which is either $\delta_{\lambda,4} - \delta_{\lambda,3}$ or $\delta_{\lambda,3} - \delta_{\lambda,4}$, or the trial 5-versus-trial 6 effect, which is either $\delta_{\lambda,6} - \delta_{\lambda,5}$ or $\delta_{\lambda,5} - \delta_{\lambda,6}$. For each of these four hypothetical between-trial effects, we subtract it from δ_2 to obtain a hypothetical posterior of δ_{GO} , and use it to compute $\text{pr}^{(h)}(\delta_{GO} > 0 \mid \text{data}) = \text{pr}(\text{survival is worse with GO compared to IA} \mid \text{data})$. For these analyses, all posteriors were computed via MCMC because, aside from the priors, no additional normality assumption is required. The four hypothetical posteriors of δ_{GO} are plotted in Figure 28.4 and summarized in the first four rows of Table 28.4.

For each of the two hypothetical trial effects with negative means, it is nearly certain that GO has a higher death rate because $\text{pr}^{(h)}(\delta_{GO} > 0 \mid \text{data}) \geq 0.99$. Even for the largest of the four hypothetical trial effects, which has mean 0.60 and sd = 0.35, the odds are 7 to 3 that GO is inferior to IA. While the actual posterior distribution of δ_{GO} cannot be determined from these data, this sensitivity analysis indicates that, for each of these four hypothetical trial effects obtained from actual trials conducted at the same institution as the GO and IA trials, it is unlikely that GO provides an improvement in survival compared to IA.

A natural question is how large the hypothetical trial effect would need to be in order for $\text{pr}^{(h)}(\delta_{GO} > 0 \mid \text{data})$ to take on specified values less than 0.50, so that GO would be preferred over IA. Rows 5–8 of Table 28.4 contain this sensitivity analysis obtained essentially by performing the previous computations in reverse. Recall that the posterior of δ_2 has mean 0.84. We equate the sd of $\delta_{\lambda,2}$ to 0.35, equate $\text{pr}^{(h)}(\delta_{GO} > 0 \mid \text{data}) = \text{pr}^{(h)}(\delta_2 - \delta_{\lambda,2} > 0 \mid \text{data})$ to a value in the range from 0.50 to

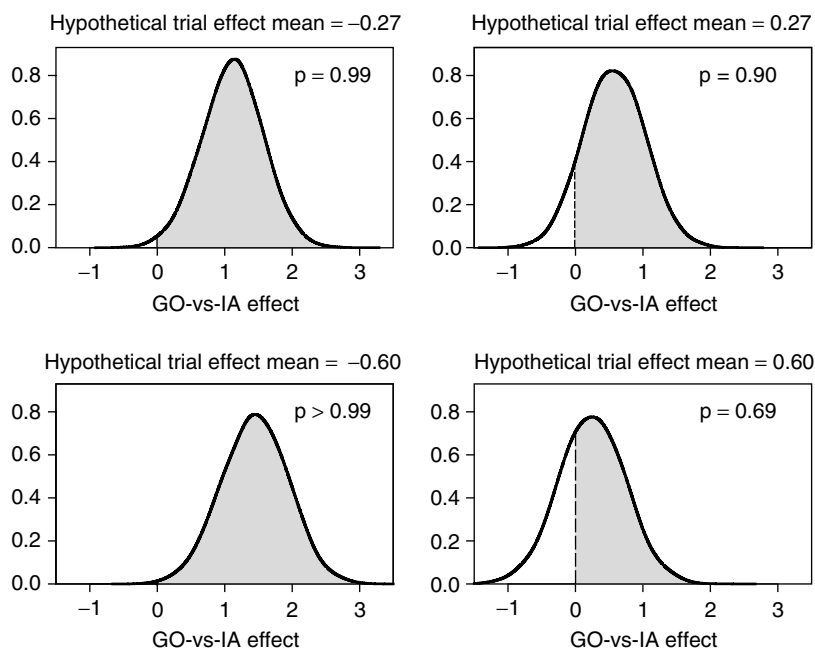


FIGURE 28.4 Hypothetical posterior distributions of the GO-versus-IA effects corresponding to each of the four actual between-trial effect distributions obtained from the FAIG and FAIGA trials. The hypothetical probability that GO is associated with worse survival than IA is $p = \text{pr}^{(h)}(\delta_{\text{GO}} > 0 \mid \text{data})$.

0.01 and, assuming normality, solve for $E^{(h)}(\delta_{\lambda})$. Table 28.4 shows that in order for the probability of GO inferiority to take on a value in the range 0.10 to 0.01 one must assume a hypothetical between-trial effect having a mean that is 2 to 3 times the largest mean between-trial effect of 0.60 previously seen.

As noted in the introduction, one may make the stronger assumption that the between-trial effects $\delta_2 = \delta_2 - \delta_1$, $\pm(\delta_3 - \delta_4)$ and $\pm(\delta_5 - \delta_6)$ are identically distributed, use the posterior of $\pm(\delta_3 - \delta_4)$ or $\pm(\delta_5 - \delta_6)$ as the prior for $\delta_{\lambda,2}$, and then integrate $\delta_{\lambda,2}$ out of $p(\delta_{\lambda,2}, \delta_{\text{GO}} \mid \text{data})$ under the assumption that $\delta_2 = \delta_{\text{GO}} + \delta_{\lambda,2}$ to obtain a posterior $p(\delta_{\text{GO}} \mid \text{data})$ of the treatment effect of interest. While computationally this is precisely what we have done in our second sensitivity analysis, an important point is that we do not assume either a prior or a posterior for $\delta_{\lambda,2}$ but rather simply assume hypothetical posteriors for $\delta_{\lambda,2}$ and assess the consequences of these assumptions.

28.3 ANALYZING CONFOUNDED COUNT DATA

28.3.1 DATA AND PROBLEM DEFINITION

The data summarized in Table 28.5 arose from two sources. The first was a single-arm clinical trial of intravenous Busulfan and Cytosin (IVBuCy) as a conditioning

TABLE 28.4

Assumed Hypothetical Trial Effect, the Corresponding Hypothetical GO-versus-IA Treatment Effect, and the Hypothetical Posterior Probability $\text{pr}^{(h)}(\delta_{\text{GO}} > 0 \mid \text{data})$ that GO is Associated with Worse Survival Compared to IA

Assumed Hypothetical Trial Effect Mean _{sd}	Corresponding Hypothetical GO Effect Mean _{sd}	Posterior Probability that GO is Inferior to IA
-0.27 _{0.30}	1.12 _{0.45}	>0.99
0.27 _{0.30}	0.57 _{0.45}	0.90
-0.60 _{0.36}	1.44 _{0.49}	>0.99
0.60 _{0.36}	0.25 _{0.49}	0.69
0.84 _{0.35}	0 _{0.48}	0.50
1.45 _{0.35}	-0.61 _{0.48}	0.10
1.62 _{0.35}	-0.78 _{0.48}	0.05
1.94 _{0.35}	-1.10 _{0.48}	0.01

TABLE 28.5

The 100-day Survival of 1,812 CML Patients for Each Confounded Preparative Regimen-Medical Center Combination

Prognostic Subgroup	# Deaths within 100 days/# Patients (%)		Posterior Probability that IVBuCy-MDACC Has Lower 100-Day Mortality than Alt-IBMTR
	IVBuCy at MDACC	Alternative Preparative Regimens at IBMTR Centers	
Chronic phase	0/17 (0)	242/1344 (18)	0.991
Accel. phase	0/25 (0)	84/335 (25)	>0.999
Blast crisis	0/5 (0)	26/86 (30)	0.945
Overall	0/47 (0)	352/1765 (20)	0.990

regimen for 47 allogeneic blood or marrow transplantation (allotx) patients with chronic myelogenous leukemia (CML). This trial was conducted at M.D. Anderson Cancer Center (MDACC) from July 1996 to October 1999. Additional details are provided by Andersson et al.¹³ The second data source was the International Bone Marrow Transplantation Registry (IBMTR), which provided cross-tabulated counts of prognostic category and the indicator of 100-day mortality for 1765 CML patients who received allotx with alternative preparative regimens (Alt), primarily Cytosan with total body irradiation or oral Busulfan and Cyclophosphamide. The three CML disease stages are chronic phase (CP), accelerated phase (AP), and blast crisis (BC). Table 28.5 illustrates the well-known fact that the probability of short-term survival in CML patients undergoing allotx decreases ordinarily with disease stage, with CP the best and BC the worst prognostic stage. While other covariates were available for the MDACC patients, after accounting for disease stage none were of any additional value for predicting the probability of 100-day survival. No patient prognostic covariates other than disease stage were available for the IBMTR patients. The data that we will utilize for comparison of the two conditioning regimens, IVBuCy and Alt, consist of 100-day mortality counts given in Table 28.5.

The main scientific difficulty with these data arises from the fact that patients were not randomized between IVBuCy and Alt. Instead, all 47 IVBuCy patients were transplanted at MDACC while all 1765 Alt patients were transplanted at IBMTR medical centers, so the two preparative regimens are confounded with the centers. We will denote these two confounded treatment-center groups as IVBuCy-MDACC for IVBuCy at MDACC and Alt-IBMTR for Alt regimens at the IBMTR centers. Within each row of [Table 28.5](#), each comparison reflects the difference between these confounded treatment-center effects rather than the IVBuCy-versus-Alt treatment effect. Thus, these preliminary comparisons are potentially misleading in that they ignore treatment-center confounding.

A visual inspection of the event rates shows the obvious and important fact that none of the IVBuCy-MDACC patients died within the first 100 days post transplant, whereas the 100-day mortality rates for the Alt-IBMTR patients varied from 18% to 30% depending on CML stage. In addition to treatment-center confounding, two other interesting aspects of these data are that all of the IVBuCy-MDACC counts are 0, and the IVBuCy-MDACC sample size of 47 is much smaller than the 1745 Alt-IBMTR sample size. In summary, the problem is to compare treatment effects between two sets of binomial samples accounting for prognostic subgroup, in the presence of treatment-center confounding, based on disproportionate sample sizes where all counts in the smaller sample are 0.

28.3.2 PROBABILITY MODEL

The following analysis is similar to that given by Thall, Champlin, and Andersson.¹⁴ Index the CML stage subgroups by $j = \text{CP, AP, BC}$, and denote the 100-day mortality probabilities in subgroup j by $\gamma_{i,j}$ for IVBuCy-MDACC patients and $\gamma_{2,j}$ for Alt-IBMTR patients. We assume *a priori* that each $\gamma_{i,j} \sim \text{iid beta}(0.5, 0.5)$ for $i = 1, 2$ and $j = 1, 2, 3$. Because the sum of the parameters of a beta distribution may be considered its effective sample size, the $\text{beta}(0.5, 0.5)$ prior contains as much information as knowing the outcome of one patient, in this case knowing whether the patient died within 100 days. Denote by $X_{i,j}$ the number of deaths and by $N_{i,j}$ the number of patients from study i and subgroup j . Then the posterior distribution of $\gamma_{i,j}$ given $X_{i,j}$ and $N_{i,j}$ is $\text{beta}(0.5 + X_{i,j}, 0.5 + N_{i,j} - X_{i,j})$, which has mean $\mu_{i,j} = (X_{i,j} + 0.5) / (N_{i,j} + 1)$, variance $\sigma_{i,j}^2 = \mu_{i,j}(1 - \mu_{i,j}) / (N_{i,j} + 2)$ and effective sample size $M_{i,j} = N_{i,j} + 1$. Observing that 0/17 IVBuCy-MDACC patients in CP died within 100 days ([Table 28.5](#)) gives a $\text{beta}(0.5, 17.5)$ posterior for $\gamma_{1,\text{CP}}$, which has mean 0.028 (sd = 0.038). Observing that 242/1344 Alt-IBMTR CP patients died within 100 days gives a $\text{beta}(242.5, 1102.5)$ posterior for $\gamma_{2,\text{CP}}$, which has mean 0.180 (sd = 0.010) reflecting both the higher observed 100-day mortality rate and the much larger sample size. The posteriors for the other prognostic groups are computed analogously. These posteriors are graphed in [Figure 28.5](#), which illustrates the fact that, because 0/47 IVBuCy-MDACC patients died within 100 days, within each prognostic subgroup and overall they had a much smaller posterior probability of 100-day mortality than the Alt-IBMTR patients.

An intuitively appealing Bayesian statistic for comparing $\gamma_{2,j}$ and $\gamma_{1,j}$ is $\text{pr}(\gamma_{2,j} > \gamma_{1,j} \mid \text{data}_j)$, the posterior probability that the 100-day mortality rate in Alt-IBMTR patients

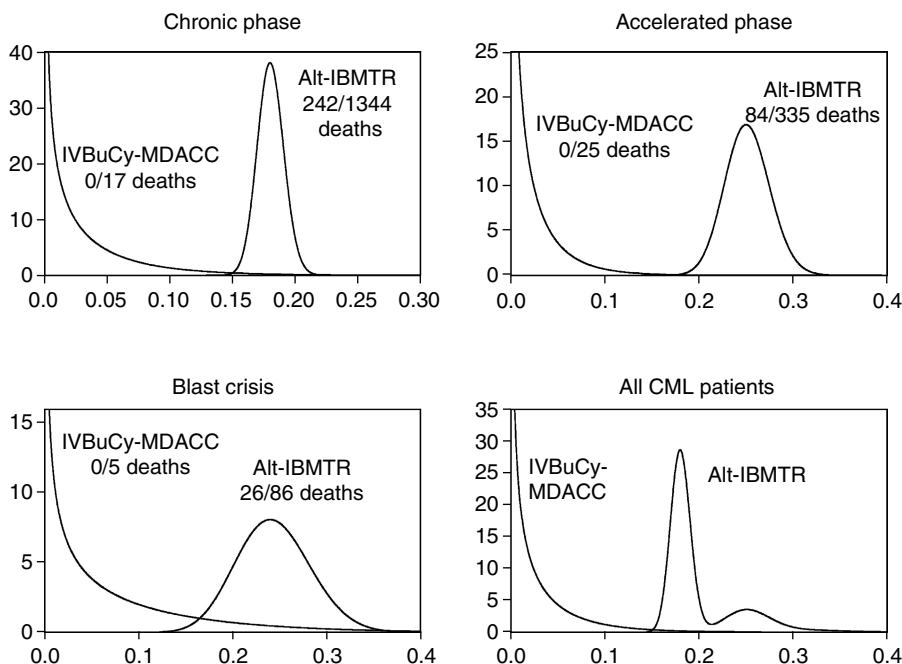


FIGURE 28.5 Posterior distributions of the IVBuCy-MDACC and Alt-IBMTR 100-day mortality probabilities within each prognostic subgroup and overall.

is higher than in IVBuCy-MDACC patients in subgroup j . This probability is 0.50 if the posteriors of $\gamma_{1,j}$ and $\gamma_{2,j}$ are identical and values greater (less) than 0.50 correspond to a lower (higher) 100-day death rate in the IVBuCy-MDACC patients. A single overall probability may be obtained by averaging the probabilities of the individual subgroups. Weighting proportional to sample size (Table 28.5), the sample proportions are $w_{CP} = 0.75$, $w_{AP} = 0.20$, and $w_{BC} = 0.05$, so the average is $\sum_{j=CP,AP,BC} w_j \text{pr}(\gamma_{2,j} > \gamma_{1,j} | \text{data}_j)$. The values of $\text{pr}(\gamma_{2,j} > \gamma_{1,j} | \text{data}_j)$ for $j = CP, AP, BC$ and the weighted average are given in the last column of Table 28.5. These indicate that, assuming uninformative priors, it is virtually certain *a posteriori* that IVBuCy-MDACC had a lower 100-day mortality probability than Alt-IBMTR (1) for the CP patients, (2) for the AP patients (3), on average across the three prognostic groups, and (4) the odds are about 17 to 1 in favor of IVBuCy-MDACC for the BC patients. The issue now is what may be said about the IVBuCy-Alt treatment effect based on these data.

28.3.3 SENSITIVITY ANALYSES OF THE TRANSPLANT DATA

Our approach is motivated in part by a desire for computational convenience, because it only requires one to compute probabilities of the form $\text{pr}(\pi_1 > \pi_2)$ for parameters π_1 and π_2 that follow beta distributions. First, consider a single disease subtype, temporarily suppress the index j , and consider the confounded effect, $\delta = \gamma_2 - \gamma_1$, of Alt-IBMTR versus IVBuCy-MDACC on 100-day mortality. Let p be the hypothetical

proportion of δ accounted for by center, for $0 \leq p \leq 1$, so that the remaining $1-p$ is due to treatment. Denote by $\theta_1(p)$ the hypothetical 100-day mortality probability of an IVBuCy patient if the center effect were identical to the effect of the IBMTR. Under the decomposition $\gamma_2 = \theta_{\text{Alt}} + \lambda_2$ and $\gamma_1 = \theta_{\text{IVBuCy}} + \lambda_1$, if a hypothetical proportion p of δ is due to $\delta_\lambda = \lambda_2 - \lambda_1$ and $1-p$ is due to $\delta_\theta = \theta_{\text{Alt}} - \theta_{\text{IVBuCy}}$, then we define $\theta_1(p) = \theta_{\text{IVBuCy}} + \lambda_2$. This implies that $\gamma_2 - \theta_1(p) = \delta_\theta$, and consequently $\text{pr}^{(h)}(\gamma_2 > \theta_1(p) \mid \text{data})$ is the hypothetical probability of interest for comparing treatments. To construct the hypothetical distribution of $\theta_1(p)$, we define the weighted average $\mu_1(p) = (1-p)\mu_1 + p\mu_2$ of the posterior means of γ_1 and γ_2 and assume that $\theta_1(p) \mid \text{data} \sim \text{beta}[\mu_1(p)M_1, \{1-\mu_1(p)\}M_1]$. This says that, for each value of p , the hypothetical posterior 100-day mortality probability of an IVBuCy patient treated under the same circumstances as an IBMTR patient follows a beta distribution with mean $\mu_1(p)$ and effective sample size the same as that of γ_1 . The probability $\text{pr}^{(h)}(\gamma_2 > \theta_1(p) \mid \text{data})$ is particularly easy to compute because γ_2 and $\theta_1(p)$ follow independent beta distributions. The decomposition $\delta = \{\gamma_2 - \theta_1(p)\} + \{\theta_1(p) - \gamma_1\} = \delta_\theta(p) + \delta_\lambda(p)$ formalizes the assumption that $\{\text{confounded effect}\} = \{\text{Alt-versus-IVBuCy treatment effect}\} + \{\text{IBMTR-versus-MDACC center effect}\}$. That is, $\delta_\theta(p) = \gamma_2 - \theta_1(p)$ is the hypothetical IVBuCy-Alt treatment effect under the assumption that $p100\%$ of δ is due to center. At one extreme, if $p = 0$ then there is no center effect, $\theta_1(p) = \theta_1(0)$ has the same mean as γ_1 , $\text{pr}^{(h)}(\gamma_2 > \theta_1(p) \mid \text{data}) = \text{pr}^{(h)}(\gamma_2 > \gamma_1 \mid \text{data})$, and $E^{(h)}(\delta) = E^{(h)}(\delta_\theta)$. If $p = 0.5$, then $\theta_1(p) = \theta_1(0.5)$ has mean $0.5\mu_1 + 0.5\mu_2$, and on average half of the observed difference is due to treatment and half to center. If $p = 1$, then $\theta_1(p) = \theta_1(1)$ has the same mean as γ_2 , all of the observed effect is due to center, there is no treatment effect, and $\text{pr}^{(h)}(\gamma_2 > \theta_1(p) \mid \text{data})$ is approximately 0.5. This probability is not exactly 0.5 because γ_2 and $\theta_1(p)$ have different variances due to the fact that we have equated the effective sample size of the distribution of $\theta_1(p)$ to that of γ_1 . Reintroducing j , our sensitivity analysis will consist of evaluating $\text{pr}^{(h)}(\delta_{\theta,j}(p) > 0 \mid \text{data}) = \text{pr}^{(h)}(\gamma_{2,j} > \theta_{1,j}(p) \mid \text{data})$ as p is varied from 0 to 1 for each $j = \text{CP, AP, BC}$, and also for the weighted average of the three subgroups.

The sensitivity analyses are summarized in Table 28.6. The first column of the table, labeled 0, gives the probability $\text{pr}^{(h)}(\gamma_2 > \gamma_1 \mid \text{data})$, within each subgroup and overall, that 100-day mortality was lower in the IVBuCy-MDACC patients compared to the Alt-IBMTR patients, thus comparing the two confounded treatment-center effects. The last four columns give $\text{pr}^{(h)}(\gamma_2 > \theta_1(p) \mid \text{data})$ for $p = 0.25, 0.50, 0.75$, and 1.00 , respectively, quantifying the hypothetical treatment effects that result from assuming that 25%, 50%, 75%, or 100% of the confounded treatment-center effect is due to center. In all three CML prognostic subgroups, even if 50% of the observed advantage is due to an intrinsic superiority of MDACC over the IBMTR centers, then the probability that IVBuCy has lower 100-day mortality compared to alternative preparative regimens still varies from 0.78 to 0.94, depending on prognostic subgroup. Only under the extreme assumption that 100% of the observed difference is due to MDACC center superiority over the IBMTR do the probabilities of IVBuCy treatment superiority drop to values near 0.50.

More generally, one may assume a probability distribution $f^{(h)}(p)$ on p and compute the average $\int \text{pr}^{(h)}(\gamma_2 > \theta_1(p) \mid \text{data}) f^{(h)}(p) dp$. This formally incorporates one's uncertainty about p into the sensitivity analysis. Each value of $\text{pr}^{(h)}(\gamma_2 > \theta_1(p) \mid \text{data})$

TABLE 28.6
Sensitivity Analyses of Center (MDACC-versus-IBMTR) Effects and Preparative Regimen (IVBuCy-versus-Alt) Effects on the Probabilities of 100-Day Mortality. Each entry is the hypothetical posterior probability, $\text{pr}^{(h)}(\gamma_2 > \theta_1(p) \mid \text{data})$, that IVBuCy has lower 100-day mortality than alternative preparative regimens

Prognostic Subgroup	Assumed Hypothetical Proportion p of the Confounded Effect that is Due to MDACC-versus-IBMTR Center Effect				
	0	0.25	0.50	0.75	1.00
Chronic phase	0.991	0.950	0.859	0.720	0.552
Accel. phase	>0.999	0.992	0.936	0.778	0.527
Blast crisis	0.945	0.875	0.779	0.665	0.543
Overall	0.990	0.954	0.871	0.729	0.546

in Table 28.6 may be regarded as a special case of the above in which $f(p)$ places probability 1 on a single value of p . This computation may be carried out for each of several hypothetical distributions on p , reflecting different opinions regarding center effects, and the sensitivity analysis would then be with regard to the different values of f . To avoid numerical integration of $\text{pr}^{(h)}(\gamma_2 > \theta_1(p) \mid \text{data})f(p)$ over p , one may approximate the integral by placing all of the probability mass of $f^{(h)}(p)$ on a few values of p . We did this using the five values $p = 0.05, 0.25, 0.50, 0.75, 0.95$. For example, if one feels it is most likely that about $p = 0.50$ of the observed effect is due to MDACC superiority but allows with some small probabilities both of the possibilities that all or none of the observed effect is due to center, then one might assume $f^{(h)}(0) = 0.05, f^{(h)}(0.25) = 0.10, f^{(h)}(0.50) = 0.70, f^{(h)}(0.75) = 0.10, f^{(h)}(1.00) = 0.05$. For this choice of $f^{(h)}(p)$, the average value of $\text{pr}^{(h)}(\gamma_2 > \theta_1(p) \mid \text{data})$ for the combined prognostic subgroups equals $0.05 \times 0.990 + 0.10 \times 0.954 + 0.70 \times 0.871 + 0.10 \times 0.729 + 0.05 \times 0.546 = 0.855$. Alternatively, the distribution $f^{(h)}(0) = 0.70, f^{(h)}(0.25) = 0.20, f^{(h)}(0.50) = 0.10, f^{(h)}(0.75) = f^{(h)}(1.00) = 0$ reflects the viewpoint that the observed effect is most likely to be due entirely to actual treatment effect, but there is still some chance that up to half of the observed effect is due to center. For this distribution, the average of $\text{pr}^{(h)}(\gamma_2 > \theta_1(p) \mid \text{data})$ is 0.971. With this more general approach, the sensitivity analysis consists of evaluating $\int \text{pr}^{(h)}(\gamma_2 > \theta_1(p) \mid \text{data})f^{(h)}(p)dp$ as a function of $f^{(h)}$ as $f^{(h)}$ is varied over a reasonable set of distributions.

An alternative approach that is very similar to assuming that $\delta = \delta_\theta(p) + \delta_\lambda(p)$ and varying p is to assume that a given number of IBMTR deaths are due to center and that the rest are due to an actual Alt-versus-IVBuCy treatment effect and to vary this assumed number. Thus, $\text{pr}^{(h)}(\gamma_2 > \theta_1(p) \mid \text{data})$ is replaced by $\text{pr}(\gamma_2 > \gamma_1 \mid \text{data}^{(h)})$. For example, in the 1344 Alt-IBMTR CP patients, as the number of deaths assumed to be due to center is varied from 0 to the observed 242, one obtains results similar to those given in Table 28.6. If one assumes that, respectively, 60 (25%), 121 (50%), or 182 (75%) of the 242 deaths are due to IBMTR-MDACC center differences, the corresponding values of $\text{pr}(\gamma_2 > \gamma_1 \mid \text{data}^{(h)})$ are 0.975, 0.928, and 0.790. This type of analysis is conceptually similar to, but not the same as, the sensitivity analysis of hypothesis tests based on attributed effects in observational data described by Rosenbaum.¹⁵

28.4 DISCUSSION

The type of Bayesian sensitivity analyses described here are no substitute for conducting comparative clinical trials correctly in the first place. However, because physicians and scientists conduct many single-arm clinical trials and other experiments, especially with phase II studies, biostatisticians are routinely confronted with the problem of comparing treatments based on data from nonrandomized studies. One solution is simply to refuse to use such data for treatment comparison and insist that a randomized trial be conducted. This leaves the problem to be solved by non-statisticians, who likely will analyze the data using statistical methods that implicitly assume the data arose from a randomized experiment. Once such results are published, the consequence is that subsequent medical practice is not unlikely to be based on fallacious conclusions. That is, ignoring the problem will not make it go away. Our proposed approach is simply to assess as honestly as possible what the distribution of a treatment effect may be under reasonable assumptions about confounding effects.

ACKNOWLEDGMENT

This research was partially supported by NIH grants R01 CA 83932 and PO1 CA-49639-12. We thank the editors for their very detailed comments on two earlier drafts of this chapter, which led to substantial improvements in the exposition.

REFERENCES

1. Fisher RA. *Design of Experiments*. Edinburgh: Oliver and Boyd, 1935.
2. Meier P. Statistics and medical experimentation. *Biometrics* 1975;31:511–529.
3. Gail MH, Wieand S, and Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984;71:431–444.
4. Pocock SJ and Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial (Corr: V32 p954–955). *Biometrics* 1975; 31:103–115.
5. Estey EH and Thall PF. New designs for phase 2 clinical trials. *Blood* 2003; 102:442–448.
6. Gehan EA and Freireich EJ. Non-randomized controls in cancer clinical trials. *New England Journal of Medicine* 1974;290:198–203.
7. Gehan EA. The evaluation of therapies: historical control studies. *Statistics in Medicine* 1984;3:315–324.
8. Begg CB and Pilote L. A model for incorporating historical controls into a meta-analysis. *Biometrics* 1991;47:899–906.
9. Li Z and Begg CB. Random effects models for combining results from controlled and uncontrolled studies in meta-analysis. *Journal of the American Statistical Association* 1994;89:1523–1527.
10. Estey EH, Thall PF, Giles F, Wang XM, Cortes JE, Beran M, Pierce SA, Thomas DA, and Kantarjian HM. Gemtuzumab ozogamycin with or without interleukin 11 in patients 65 years of age or older with untreated acute myeloid leukemia and high-risk

- myelodysplastic syndrome: comparison with idarubicin + continuous-infusion high-dose cytosine arabinoside. *Blood* 2002;99:4343–4349.
11. Kaplan EL and Meier P. Nonparametric estimator from incomplete observations. *Journal of the American Statistical Association* 1958;53:457–481.
 12. Brooks SP and Gelman A. (1998) Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 1998; 7:434–455.
 13. Andersson BS, Thall PF, Madden T, Couriel D, Wang Z, Tran HT, Anderlini P, deLima M, Gajewski J, and Champlin RE. Busulfan systemic exposure relative to regimen-related toxicity and acute graft vs. host disease: defining a therapeutic window for IV BuCy2 in chronic myelogenous leukemia. *Biology of Blood and Marrow Transplantation* 2002;8:477–485.
 14. Thall PF, Champlin RE, and Andersson BS. Comparison of 100-day mortality rates associated with IV busulfan and cyclophosphamide versus other preparative regimens in allogeneic bone marrow transplantation for chronic myelogenous leukaemia: Bayesian sensitivity analyses of confounded treatment and center effects. *Bone Marrow Transplantation* 2004;33:1191–1199.
 15. Rosenbaum, PR. Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika* 2001;88:219–231.

29 Interpreting Longitudinal Studies of QOL with Nonignorable Dropout

Diane L. Fairclough

CONTENTS

29.1	Introduction	541
29.2	Why Do Missing Data Matter?	542
29.3	What Is the Question?	543
29.4	An Example	544
29.5	Analysis of Studies with Nonignorable Dropout.....	544
29.5.1	Missing Data Associated with Death	546
29.5.2	Analysis of Complete Cases	546
29.5.3	Last Observation Carried Forward.....	547
29.5.4	Maximum Likelihood Analysis and Multiple Imputation of All Available Data	547
29.5.5	Shared-Parameter Model (Joint Mixed-Effects Model).....	549
29.5.6	Pattern Mixture Models	550
29.5.7	Selection Models	551
29.6	Sensitivity Analyses	552
29.7	Summary	554
	References	554

29.1 INTRODUCTION

Missing data are inevitable in any longitudinal study in which over an extended period of time patients experience morbidity or mortality due to disease or its treatment. This is especially true for trials involving patients with cancer. While discontinuation of treatment at the patient's request is rare, therapy is often discontinued as a result of excessive toxicity, disease progression, and death. When the assessment of patient reported outcomes, such as symptoms and quality of life (QOL), is stopped after the discontinuation of treatment, it is highly likely that the outcome differs between those patients who have and have not been measured. Specifically, the causes of dropout will result in negative outcomes such as increased symptom

burden that will in turn impact the patient's quality of life. We cannot ignore this missing data if we wish to obtain unbiased estimates of the longitudinal changes. The focus of this chapter is the interpretation of the results from various analytic strategies of studies with nonignorable dropout. Some missing data may also be unrelated to the patient's QOL resulting from administrative problems such as staff forgetting to give the forms to the patient during a very busy time. This unrelated missing data is preventable and should be minimized in a well-conducted trial.

Missing data can be classified into one of three groups.¹ Briefly, if the reasons for missing assessments are completely unrelated to the patient's QOL, then we consider the data to be missing completely at random (MCAR). If the probability of a missing assessment only depends on the previously observed measures of QOL and other explanatory factors, we consider the data to be missing at random (MAR). Finally, missing data are considered to be missing not at random (MNAR), or nonignorable, if the probability that an observation is missing depends on the value of the missing observation. This would be typical of dropout that is associated with excessive toxicity or disease progression. In this chapter, I will use the term *nonignorable* interchangeably with MNAR without necessarily implying the stricter definition that includes the distinctness of parameters of the missing data and outcome models.

29.2 WHY DO MISSING DATA MATTER?

Missing data are problematic for several reasons. First is the loss of power to detect change over time or differences between groups as a result of a reduced number of observations. However, in many large trials, the sample size has been based on other clinical endpoints, and the power to detect meaningful differences in QOL measures is generally adequate. If not, increasing the sample size of the trial may be feasible depending on patient and economic resources.

Second is the potential for bias of the estimates as a result of nonignorable missing data. Patients who are experiencing a negative impact of the disease or therapy on their lives may be less likely to complete the QOL assessments. Olschewski et al.² describe a worst-case scenario where dropout in one arm occurs because the treatment is ineffective and the patient's health status deteriorates. In contrast, dropout in the other arm is minimal despite some toxicity because the treatment is effective. They point out that the ineffective treatment may appear better because the QOL is being reported only by a selected group of subjects who have not yet deteriorated because of their disease. This potential for bias can be illustrated in an observational study of 68 patients with metastatic or progressive disease who have failed prior therapy and are currently receiving palliative care.³ These patients completed an assessment of QOL every 3–6 weeks. Figure 29.1 shows estimates of the mean trajectories under the different assumptions for the dropout process. When the data are analyzed using a method that assumes the data are ignorable (maximum likelihood estimation of all available data) there is only a small insignificant decrease in the estimates of QOL over time. However, when the observations are analyzed using two approaches that consider the dropout to be nonignorable (shared parameter and pattern mixture models), the estimates exhibit a rapid rate of decline in QOL over time.

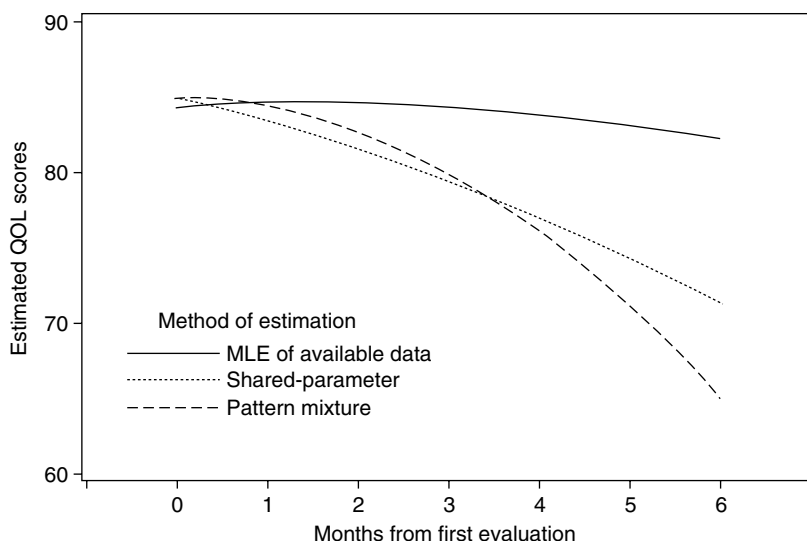


FIGURE 29.1 Estimated changes in 68 terminal cancer patients using one method of analysis that assumes dropout can be ignored (ML estimation of all available data) and two methods that assume dropout is not ignorable (pattern mixture and shared parameter model).

29.3 WHAT IS THE QUESTION?

Perhaps the most critical issue that impacts the interpretation of results from a longitudinal study of QOL in patients with cancer is the group for which we wish to make inferences. Does it include all patients who are started on a treatment regimen, those who remain on treatment, or only survivors? This is a critical distinction. For example, are we trying to answer the question, “Which treatment choice, on the average, will result in the best quality of life?” or “Among those who will be long term survivors, which group will experience better QOL?” In the first setting, we may consider QOL as a primary outcome where the decisions about alternative treatment regimens will be made on the basis of QOL. This would be particularly relevant if survival is similar among the treatment options. This question would be relevant across the entire prognostic spectrum. In most men diagnosed with prostate cancer, the survival advantages of treatment are minimal. Further, the long-term impact of the various interventions are quite different. Similarly, for patients with very advanced disease, the differences in survival among different treatment options are often in the range of several months. Patients may be more interested in the quality of the time they have rather than the quantity. In contrast, when there are substantial differences in survival, the role of quality of life measures is likely to be quite different. QOL measures may be used to identify treatable side-effects of therapy or to identify a specific component of the survivor’s well-being. For example, Moinpour et al.⁴ observed that patients with advanced prostate cancer receiving flutamide reported worse emotional functioning than those receiving placebo. Considerable caution should be employed in comparing treatment groups when

analyses are limited to patients who remain on treatment or who are survivors. All the advantages of randomization are lost, and the two groups may be different (or fail to be different) due to selection bias. For example, subjects on a particular treatment may be experiencing more severe side effects and poorer QOL outcomes, and as a result they discontinue the treatment. If the analysis is limited to patients who remain on the study, data are only obtained from subjects who are tolerating the therapy, and thus the two arms will appear to be similar.

29.4 AN EXAMPLE

To illustrate some of the issues associated with the interpretation of a trial with non-ignorable dropout, consider a trial designed to evaluate the effectiveness of two paclitaxel–cisplatin regimens with a traditional etoposide–cisplatin regimen for the treatment of stage IIIB–IV non-small-cell lung cancer (NCSLC).⁵ In addition to traditional endpoints, such as time to disease progression and survival, QOL was included in this study. The Functional Assessment of Cancer Therapy, Lung Cancer Version 2 (FACT-L)⁶ was used to measure QOL prior to therapy and approximately 6, 12, and 26 weeks after the initiation of therapy. This instrument is designed to assess four components of well-being (physical, functional, social, and emotional) as well as seven lung cancer symptoms; scores are presented on a scale of 0 to 100 with higher values indicating a better outcome. A total of 1402 QOL assessments were obtained from 575 patients. This represents 94, 72, 60, and 50% of the surviving patients at each of the four planned assessment times. Approximately a third of the patients died within 26 weeks. When the subjects are divided into groups based on the timing of their last assessments, each group contains approximately a quarter of the patients. The mean scores for four of the subscales are presented by time of the last assessment in [Figure 29.2](#). Note that for both physical and functional well-being, the baseline scores are lowest for those who drop out earliest, and the scores start dropping prior to dropout. This clearly rules out an assumption that dropout is MCAR for these outcomes, as dropout depends on previously observed measures of physical and functional well-being. In contrast, the observed scores before dropout for emotional and social well-being do not show as much difference between the groups, and it is not obvious that dropout is associated with these outcomes.

29.5 ANALYSIS OF STUDIES WITH NONIGNORABLE DROPOUT

As discussed in the chapter by Troxel and Moinpour, it is not possible to statistically test for the presence of nonignorable dropout. The information that is needed to derive a test is exactly what is missing. Often the reasons for dropout and clinical observation lead us to believe that we must consider dropout to be nonignorable. This is especially true in trials involving cancer patients, where the primary causes of dropout are excessive toxicity and disease progression. Because we lack the data that would allow us to differentiate among various dropout mechanisms, the general recommendation is to perform analyses using a number of approaches. Many of the

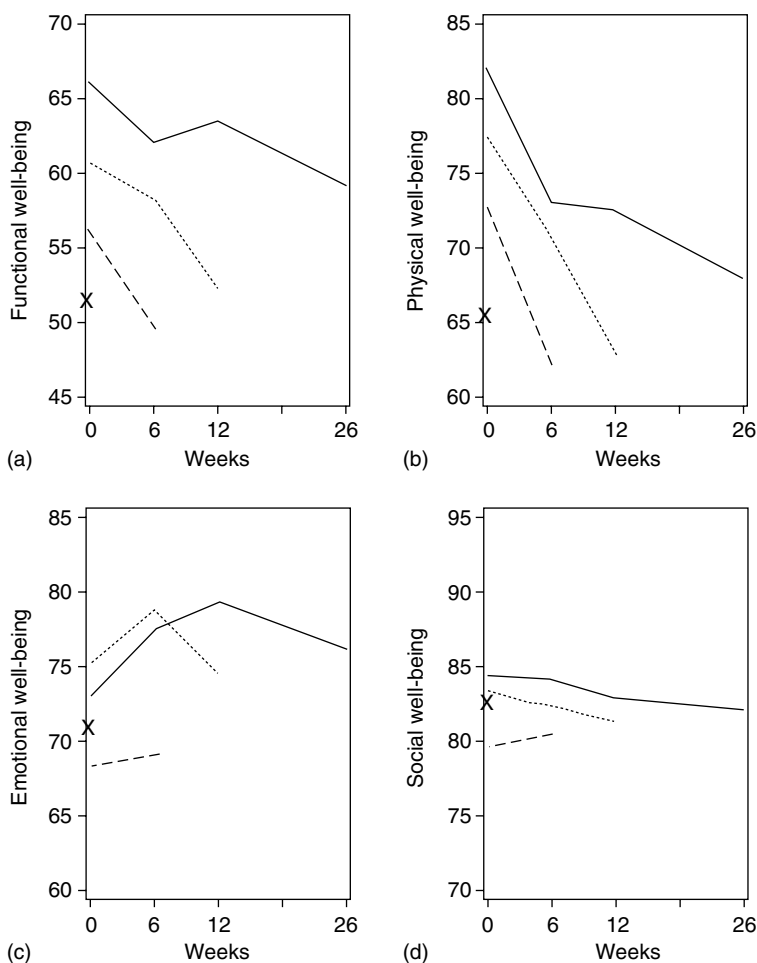


FIGURE 29.2 Means of the observed scores from four scales of the Functional Assessment of Cancer Therapy (FACT-L) in patients with NSCLC grouped by the time of their last assessment.

possibilities are described in the chapter by Troxel and Moinpour. When the conclusions are consistent across all reasonable approaches, we gain confidence that the conclusions are not sensitive to our assumptions. When the results are inconsistent, our ability to interpret the results becomes more difficult. In the remainder of this chapter, I am going to discuss a number of popular approaches to the analysis of longitudinal studies. The differences between the approaches become clear when we realize that all the methods impute the missing values. The imputation is explicit in some cases and implicit in others. Graphical display of both the observed and imputed data by pattern of dropout allows us to assess the reasonableness of the underlying assumptions. When the results are consistent with clinical observation or a value assessment that is justifiable, then our confidence in the overall results increases.

29.5.1 MISSING DATA ASSOCIATED WITH DEATH

There are numerous viewpoints about strategies for handling missing data that occur after death. Some individuals argue that it does not make sense to impute missing QOL scores that occur after death because one cannot assign value to life when life no longer exists. If one adopts that philosophy, one severely limits the questions that can be answered to those relevant to populations of survivors. Comparisons between different treatment regimens may not be possible if selection bias exists. One additional philosophic question is whether the quality of one's life can be worse than death. This may be less of an issue for patients with cancer than those with neurocognitive diseases.

Whether or not the analyst explicitly imputes missing QOL data after death, some form of imputation occurs in the analysis that influences the results and their interpretation and thus cannot be ignored. When the measure of QOL is a utility score elicited using time-trade-off, standard gamble, or another related method, the scale is often explicitly anchored with 0 representing the health state of death and 1 representing perfect health. Most other scales do not have this anchoring, so a large number of strategies have been proposed. In addition to the analytic methods that will be described in the later sections of this chapter, there have been a number of imputation schemes that view death as the worst possible outcome for patients with cancer. Some of the options include assigning the worst possible score, assigning the worst observed score, and last value carried forward. Diehr et al.⁷ explored a number of methods, including ignoring deaths, assigning an arbitrary value (0 on a 0–100 point scale), and probabilities of remaining healthy or alive in 2 years. They note that strategies that gave less influence to death tended to show more favorable changes in health status over time and therefore favored the group that had more deaths. The strategies that gave more influence to death favored the group with fewer deaths.

Another group of strategies is based on the idea that we can define groups whose ordering reflects an ordering of outcomes. Gould⁸ describes an approach where if there is adequate documentation concerning the reasons for missing assessments, it may be possible to determine a reasonable ordering (or ranking) of QOL among the subjects. For example, it would be reasonable to assign patients who withdraw because of disease progression, excessive toxicity, or death a rank that is lower than that observed for patients remaining on the study. The advantage of this approach is that we do not have to impute the specific value. Heyting et al.⁹ identify some limitations, including multiple reasons for drop-out that are not clearly ordered. Ware et al.¹⁰ in a study of 4-year health outcomes as part of the Medical Outcomes Study, defined three outcome groups: improved, stable, and declined. Patients who died during the interval were classified as having declined.

29.5.2 ANALYSIS OF COMPLETE CASES

Multivariate analysis of variance (MANOVA) or growth curve models that only include data from patients who have completed *all* of the scheduled assessments (complete cases) is the least desirable approach and increasingly less common. If the proportion of subjects with any missing assessments is very small (<5% of the cases), these methods may be reasonable. However, in studies of patients with cancer

who are experiencing morbidity and/or mortality, these methods could easily exclude more than half the subjects from the analysis, yielding results applicable only to a small proportion of the subjects. In the NSCLC study, only roughly one-fourth of the subjects had complete data. When this approach is taken, we are assuming that the subjects who are excluded have the same distribution of assessments (same mean and variance) as the subjects with complete data. When this is not true, as in our example of physical and functional well-being of NSCLC patients (Figure 29.2), the results are likely to present an overoptimistic view.

29.5.3 LAST OBSERVATION CARRIED FORWARD

Another popular approach is to use the last observation (or value) carried forward (LOCF or LVCF), where the patient's last available assessment is substituted for each of the missing assessments. This approach has limited utility⁹ and should be employed with great caution. For example, in a study where QOL is decreasing over time, a treatment with early dropout could possibly look better than a treatment where tolerance to the therapy was better. If employed, the assumptions should be explicitly stated and examined. If it is likely that the values for those dropping out due to lack of efficacy or to side-effects will be lower than for those remaining on the study, LOCF is a poor choice. If we examine the functional and physical well-being scores for the NSCLC patients displayed in Figure 29.2, carrying the last value forward for subjects who drop out between the first and second assessments could be visualized by drawing a horizontal line from the first assessment. The interpretation is that the average scores in these subjects with early dropout would be approximately the same as for those who will remain in the study long enough to complete the 6 or 12 week assessments. Careful examination of the reasons for discontinuation will allow us to decide if this assumption is reasonable or interpretable.

29.5.4 MAXIMUM LIKELIHOOD ANALYSIS AND MULTIPLE IMPUTATION OF ALL AVAILABLE DATA

The assumption that data are MCAR, required in the analysis of complete cases, can be relaxed to the assumption that data are missing at random (MAR)¹ or ignorable.¹¹ Likelihood-based methods such as the EM algorithm,^{12,13} Newton–Raphson, or multiple imputation using Markov chain Monte Carlo procedures¹⁴ use all the available data and result in unbiased estimates when the data are MAR, but they will still be biased if the dropout is nonignorable. In this situation, the probability that an observation is missing may depend on observed data and covariates but still must be independent of the value of the missing observation(s) after adjustment for the observed data and covariates. The implicit imputation of the missing values is best understood if one considers the EM algorithm that is often used to obtain the estimates. In the E-step, we estimate the values of a sufficient statistic conditional on the observed data and the current estimates of the parameters; the expected values of the missing observations are the best linear unbiased predictors (BLUP). Figure 29.3 displays the means of the observed values (solid lines), and the imputed values (dashed lines) as a function of the time of dropout for functional and emotional well-being scores in the NSCLC study.

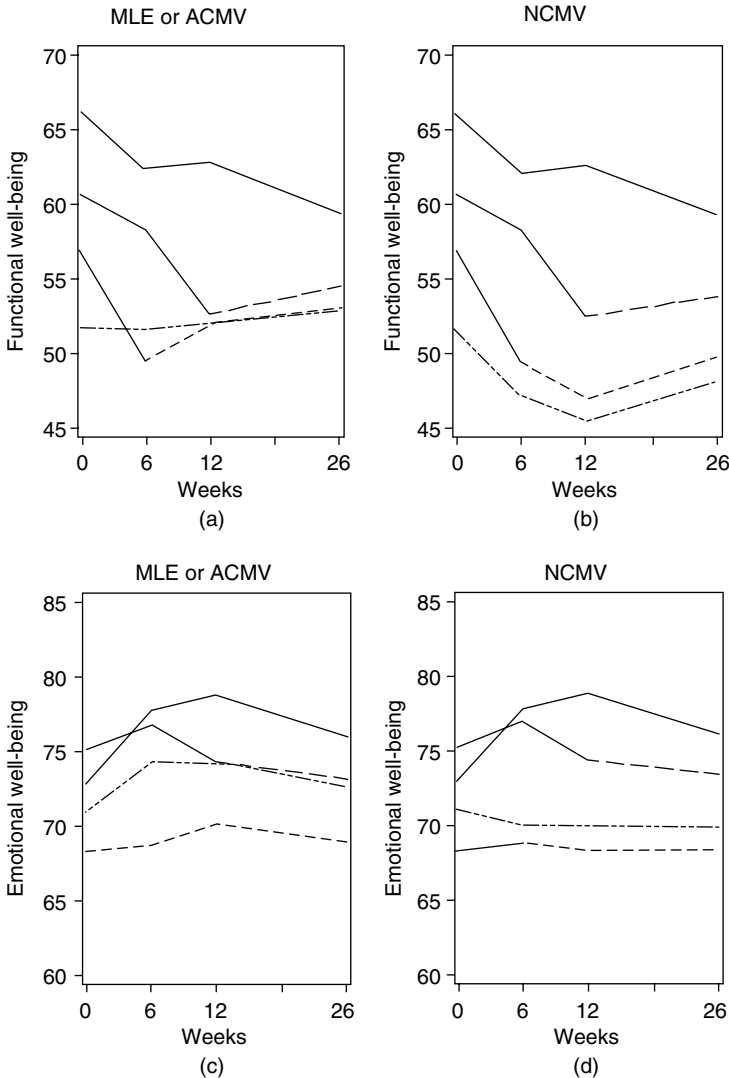


FIGURE 29.3 Means of observed (solid) and imputed (dashed) scores for functional and emotional well-being in patients with NSCLC grouped by the time of their last assessment. Imputed scores were estimated under the available case missing value (ACMV) and nearest case missing value restrictions (NCMV). The ACMV restriction is equivalent to maximum likelihood estimation (MLE) when dropout is monotone.

We note that the estimates of the missing values in the figures labeled MLE or ACMV move toward the observed data, rising after the last observed measurement. As this is not consistent with clinical observation of the functional well-being of these patients, it suggests that MLE of available data will result in biased estimates of some of the components of QOL in this setting.

29.5.5 SHARED-PARAMETER MODEL (JOINT MIXED-EFFECTS MODEL)

An example of a shared parameter model is the proposed extension of the random effects or two-stage mixed-effects model by Schluchter¹⁵ and DeGruttola and Tu.¹⁶ In this joint model, both the longitudinal QOL scores and the time of censoring (or death) are incorporated by allowing the time of censoring to be correlated with the random effects of the longitudinal model for QOL.

Returning to the NSCLC study, in Figure 29.2 we observed that subjects with a longer period of QOL assessment as a group had both higher initial scores and more positive slopes for the measures of functional and physical well-being. There was also significant variation among subjects in the random effects for the intercept and slope. This suggests that it might be reasonable to assume that the baseline values and rate of decline in each individual's physical and functional well-being over time are correlated with the time of dropout, disease progression, or death in these patients. Using this model, the estimated correlations of the random intercept and slope terms with the log of the survival times were 0.42 and 0.80 for functional well-being and 0.40 and 0.88 for physical being. Figure 29.4 plots estimated trajectories determined by the mean of random effects stratified by survival time. From this plot it is clear that patients who died earlier had both lower initial values and a more rapid rate of decline in their scores, consistent with the patterns displayed in Figure 29.2. If the extrapolation of these lines past the point of death represents a reasonable penalty for early death, then this approach will be useful for the comparison of these outcomes over the period of the study (6 months in the NSCLC example). When the overall results from the shared parameter model are compared to those obtained from MLE of all available data, the rate of decline increased from -0.8 to -2.1

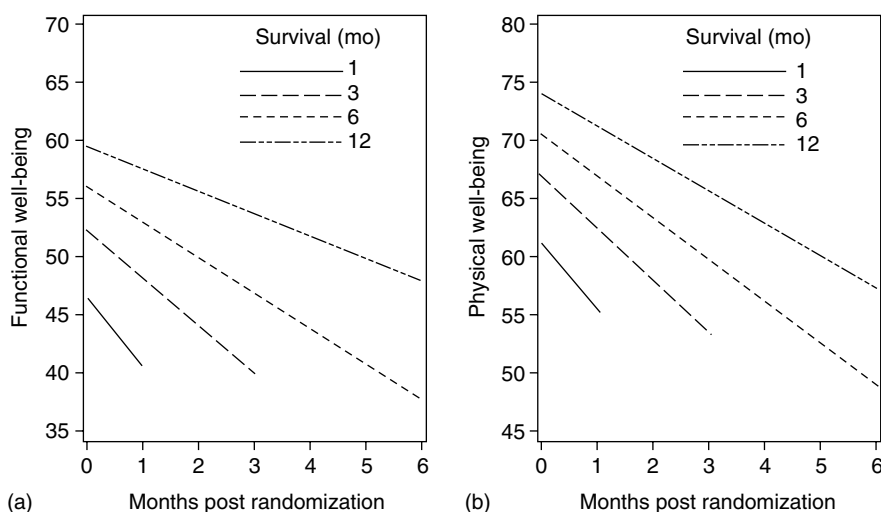


FIGURE 29.4 Conditional estimates of change of functional and physical well-being in patients with NSCLC using a shared parameter model (joint model of QOL scores with log of survival) as a function of survival.

points/month for functional well-being and from -1.9 to -2.9 points/month for physical well-being in the MLE and shared parameter models, respectively.

A couple of observations suggest the settings where this may be a useful model for outcomes for patients with advanced cancer. The first is that the correlation between survival and the rate of change in measure of symptoms, as well as physical and functional well-being, increases as the prognosis of the patients decreases. The second is that survival tends to have stronger correlations with these measures than time of dropout or time of disease progression. My speculation is that disease progression is generally measured radiographically or as a function of a biological marker; symptomatic progression of the disease and the impact on measures of QOL occurs later in time and is more closely associated with the length of survival. Finally, while there is considerable variation among individuals in their measures of emotional and social well-being, scores tend to remain stable over time within individuals, and there is minimal variation in the rates of change among individuals. As it is hard to identify correlations with something that has minimal variations, this shared parameter model is not as useful for these outcomes.

29.5.6 PATTERN MIXTURE MODELS

The basic concept behind the pattern mixture models described by Little¹⁷ is that the distribution of the measures of QOL, Y_i , may differ across the K different missing data patterns, having different means, $\mu^{(k)}$, and variance, $\Sigma^{(k)}$. For example, patients who die earlier (with missing data due to death) may have lower QOL scores (different means) and also may have more or less variability in their scores (different variance) than patients who survive longer. The true distribution of the measures of QOL for the entire group of patients will be a mixture of the distributions of each of the K groups of patients. The general method of analysis is to stratify the patients by the missing data patterns. Then the mean and variance parameters ($\hat{\mu}^{(k)}, \hat{\Sigma}^{(k)}$) are estimated within each stratum. Weights are determined by the number of subjects within each of the K missing data patterns ($w_k = n_k/N$). The population estimates are the weighted average of the estimates from the K missing data patterns ($\hat{\mu} = \Sigma w_k \hat{\mu}^{(k)}$). The advantage of these models is that one does not have to define a model for the missing data mechanism. One of the difficulties of this approach is the large number of missing data patterns that occur in actual studies where there may only be a few subjects with some patterns. But more critically, for all but the one pattern where there are no missing observations, the model may be underidentified. Specifically, it may not be possible to estimate all the parameters of the model without making additional assumptions. The most obvious example of an underidentified model occurs for the pattern where all the observations are missing. The assumptions or restrictions that allow for estimation of all the model parameters cannot be validated because they depend on the values of the missing data. This is balanced by the ease of graphically examining the implications of those restrictions.

A wide range of restrictions have been applied to pattern mixture models.^{18,19} The first group is applied to repeated measures models and includes complete case missing value (CCMV), available case missing value (ACMV), and neighboring case missing value (NCMV) restrictions. In the application of the CCMV restriction,

the implication is that the missing data are ignorable conditional on the observed data from patients with complete data and one of the missing data patterns. Too many studies of cancer patients have noted that this is not the case.

The NCMV restriction represents the other extreme; we are assuming that the missing data are ignorable among subjects with the closest patterns. For example, data from subjects who drop out after the second assessment would be used to estimate the missing values for the second assessment among patients who drop out after the first assessment. In most cancer trials, this would seem to be a more reasonable assumption. This is illustrated in [Figure 29.3](#), where the imputed missing values of functional well-being in the NCSLC study tend to drop after the last observed assessment. Unfortunately, we are ultimately required to use data from those who complete all assessments to impute the missing data in the final assessment. The values that are imputed at the final assessment rise as a result. This counterintuitive observation suggests that this restriction will result in an overly optimistic estimate of the functional well-being of these patients at 26 weeks.

The ACMV restriction uses all available cases to impute missing values; if the missing data pattern is strictly monotone (subjects are never observed after their first missing assessment), the results are identical to those obtained for MLE of all available data.

The second group of restrictions that are most commonly imposed are extrapolations of growth-curve models, generally polynomial models.²⁰ As previously mentioned, the framework of pattern mixture models allows the analyst to examine the results of these extrapolations for each of the patterns of missing data. However, there is still the challenge of specifying the necessary restrictions. This is extremely difficult to do before examining the data, which leaves the analysis vulnerable to criticism. The NSCLC study provides an example where a pattern mixture model is easiest to implement because the changes over time appear to be roughly linear during the periods where the subjects are observed, thus simplifying the polynomial model. However, we still need to believe that it is reasonable to extrapolate the linear functions, and we need to make a decision about how to specify the linear function for subjects who were only observed at baseline.

29.5.7 SELECTION MODELS

In selection models, a statistical model is specified for the missing data mechanism. In addition to adjusting the estimates for the missing data, these models allow the investigator to make inferences about the relationship of QOL and other explanatory factors causing missing observations. This might be particularly interesting, for example, if death or disease progression was the cause of dropout. Selection models for the analysis of studies with nonignorable missing data include models where the change in QOL over time is functionally related to the time of death, disease progression, or study dropout. Wu and Carroll²¹ proposed a probit model for the probability of dropout, which depends on the linear change in QOL over time. The most well-known of the selection models was proposed by Diggle and Kenward.²² They proposed that the probability of dropout could be a function of subject characteristics (covariates), the past history (previous observations), and the current value of the

outcome. The ability to estimate the relation of dropout to the value of the QOL measure at the time of dropout depends entirely on the distribution that is specified for the outcome score. While initially appearing to have great promise, this model has been extensively criticized because of the sensitivity of the results to this assumption. This is particularly relevant for QOL measures. In many studies, QOL scores are skewed with longer tails for lower scores. If subjects with poorer QOL are more likely to be missing, the nonmissing scores will appear to have a roughly normal distribution. If the scores are assumed to be normal, the most common assumption, they may appear to be random when in fact they are not.

29.6 SENSITIVITY ANALYSES

Given the numerous potential methods of analysis, how do we choose between different strategies? In some cases, we will have information such as the reason for missing assessments or a clearly defined objective that will determine the best approach. But in general, while certain approaches may be eliminated from consideration, we will be left with several possibilities. A sensitivity analysis in which the effect of the different methods of analysis is examined may be informative. Graphical examination of the values that are explicitly or implicitly imputed as has been illustrated in this chapter will be a useful step when determining which results are interpretable and how they are interpreted.

There are two likely outcomes of the sensitivity analysis. The first is that the conclusions are consistent regardless of the approach. An example of this is a sensitivity analysis of four methods of handling missing data in a clinical trial of two adjuvant therapies for breast cancer.²³ In both arms, 9% of the patients were missing the assessment that occurred during therapy. The endpoint was the change in QOL (during-before therapy) as measured by the Breast Chemotherapy Questionnaire (BCQ).²⁴ In a nonparametric two-sample comparison (Wilcoxon rank-sum test), the four strategies considered included 1) analysis of available data, 2) assigning the lowest rank when missing data occurred in patients discontinuing treatment because of toxicity or relapse, 3) simple imputation of the missing data using regression, and 4) assigning the lowest rank when missing data occurred regardless of the reason.²⁵ Regardless of the approach taken, medians and interquartile ranges were similar, and the treatment differences were statistically significant. The alternative outcome of a sensitivity analysis is that the conclusions are dependent on the method of analysis or the summary measure selected. When this occurs, the methods should be examined to ascertain the reason for the discrepancy. For example, [Figure 29.1](#) illustrates the dramatic differences in the estimates of change over time in the study of the 68 terminal cancer patients between the analysis that assumed the missing data was missing at random and the two that considered nonrandom patterns.

[Figure 29.5](#) shows the estimates of change over time in physical and functional well-being of NSCLC patients are sensitive to the method of analysis. As we progress from analyses that assume dropout is completely random (complete case analysis), to the two that assume the data are MAR (ML repeated measures and ML mixed effects model), to the two methods that consider the dropout to be nonignorable, the estimates at baseline decline and the rate of decrease in scores declines more rapidly. In contrast, measures of

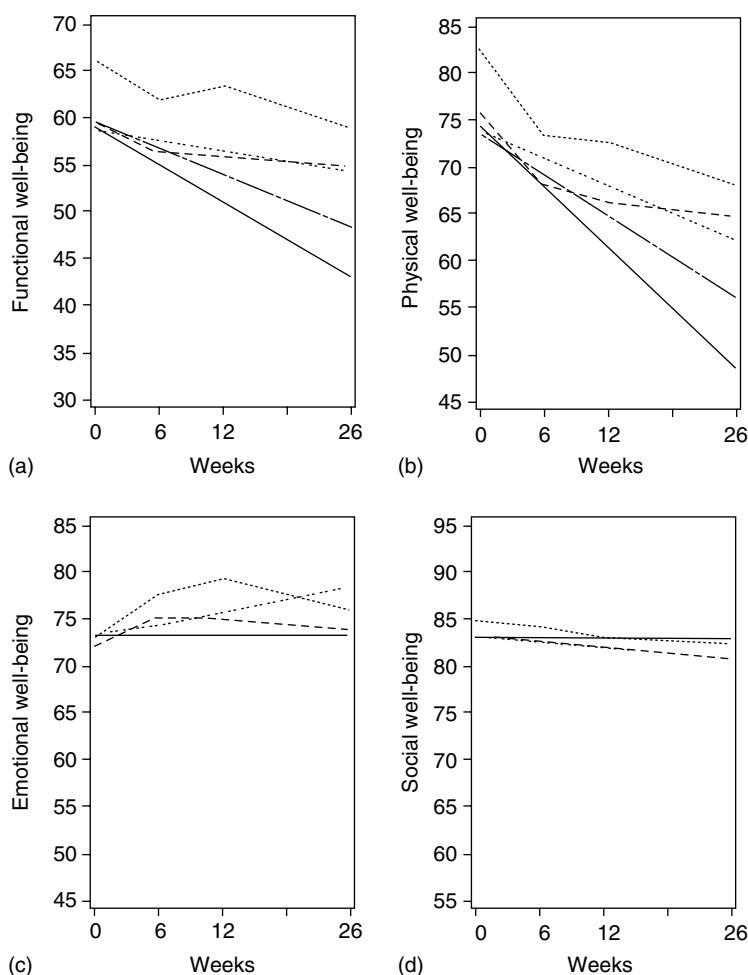


FIGURE 29.5 Estimates of change for four scales of the FACIL-L in patients with NSCLC. Five methods of analysis are displayed: (· · · · ·) complete case; (— — — —) ML of repeated measures; (— — — —) ML mixed effects model; (- - - -) pattern mixture with linear extrapolation; (- · · · · ·) Shared parameter model (not displayed for emotional and social well-being).

emotional and social well-being are relatively insensitive to the method of analysis. How are treatment comparisons affected? In contrast to estimates of change over time, the results of between group comparisons remained remarkably consistent across different analysis models²⁶ including an analysis of complete cases (MCAR), all available data (MAR), and a shared-parameter model (NMAR). In trials of patients with advanced disease, one is likely to observe a high rate of nonignorable dropout and should expect estimates of change within groups to be very sensitive to the chosen method of analysis. In contrast, between group comparisons are likely to be much less sensitive, especially when the rates and reasons for dropout are similar across treatment groups.

29.7 SUMMARY

The impact of missing data in QOL studies should be carefully evaluated. Nonignorable missing data are likely in QOL studies where there is dropout due to toxicity, disease progression, or even therapeutic effectiveness. Studies with this type of missing data are also the most difficult to analyze. The primary reason is that there are numerous potential models, and it is impossible to verify statistically the “correctness” of any model because the data required to distinguish between models are missing. In practice, current application of methods (selection, pattern mixture, and shared-parameter models) for nonignorable missing data is limited by several factors. The first is the large number of subjects that will be required to distinguish between alternate models. The second is the restriction of some assumptions such as linear changes over time and the inability of many techniques to deal with both dropout and intermittent missing data. In addition, the sophisticated programming required for these methods and the lack of generally available software are barriers to implementation. However, the most significant barrier may be the difficulty of presenting these complicated models in a manner that is readily interpretable in the clinical literature.

REFERENCES

1. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, 1987.
2. Olschewski M., Schulgen G., Schumacher M., Altman DG. Quality of life assessment in clinical cancer research. *Br J Cancer* 1994; 70:1–5.
3. Hwang SS, Chang VT, Fairclough DF, Hoch K, Latorre M, Corpion C. Longitudinal measurement for quality of life and symptoms in terminal cancer patients. *Proceeding of the Am. Soc. for Clinical Oncol.* 1996; abstract.
4. Moinpour CM, Savage MJ, Troxel A, Lovato LC, Eisenberger M, Veith RW, et al. Quality of life in advanced prostate cancer: results of a randomized therapeutic trial. *J Natl Cancer Inst* 1998; 90:1537–44.
5. Bonomi P, Kim KM, Fairclough DL, Cella D, Kurgler J, Rowinsky E, et al. Comparison of survival and quality of life in advanced non-small-cell lung cancer patients treated with two dose levels of paclitaxel combined with cisplatin versus etoposide–cisplatin: results of an Eastern Cooperative Group Trial. *J of Clin Oncol* 2000; 18:623–31.
6. Cella DF, Bonomi AE, Lloyd SR, Tulskey DS, Kaplan E, Bonomi P. Reliability and validity of the Functional Assessment of Cancer Therapy — Lung (FACT-L) quality of life instrument. *Lung Cancer* 1995; 12:199–220.
7. Diehr P, Patrick D, Hedrick S, Rothman M, Grembowski D, Raghunathan TE, et al. Including deaths when measuring health status over time. *Med Care* 1995; 33:AS164–72.
8. Gould AL. A new approach to the analysis of clinical drug trials with withdrawals. *Biometrics* 1980; 36:721–7.
9. Heyting A, Tolboom JT, Essers JG. Statistical handling of drop-outs in longitudinal clinical trials. *Stat Med* 1992; 11:2043–61.
10. Ware JE, Bayless MS, Rogers WH, Kosinski M, Tarlov AR. Differences in 4-year health outcomes for elderly and poor, chronically ill patients treated in HMO and fee-

- for-service systems: results from the Medical Outcomes Study. *JAMA* 1996; 276:1039–47.
11. Laird NM. Missing data in longitudinal studies. *Stat Med* 1988; 7:305–5.
 12. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM Algorithm. *J R Stat Soc (B)* 1977; 39:1–38.
 13. Jennrich RI, Schluchter MD. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* 1986; 42:805–20.
 14. Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Beh Res* 1998; 33:454–71.
 15. Schluchter MD. Methods for the analysis of informatively censored longitudinal data. *Stat Med* 1992; 11:1861–70.
 16. DeGruttola V, Tu XM. Modeling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics* 1994; 50:1003–14.
 17. Little RJA. A class of pattern-mixture models for normal incomplete data. *Biometrika* 1994; 81:471–83.
 18. Little RJA. Pattern-mixture models for multivariate incomplete data. *J Am Stat Assoc* 1993; 88:125–34.
 19. Curran D. Analysis of Incomplete Longitudinal Quality of Life Data. Ph.D. dissertation. Limburgs Universitair Centrum, 2000.
 20. Fairclough DL. Pattern mixture models. In: *Design and Analysis of Quality of Life Studies in Clinical Trials*. Boca Raton: Chapman & Hall/CRC, 2002:153–81.
 21. Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 1988; 44:175–88.
 22. Diggle P, Kenward MG. Informative drop-out in longitudinal data analysis. *Appl Stat* 1994; 43:49–93.
 23. Fetting JH, Gray R, Fairclough DL, Smith TJ, Margolin KA, Citron ML, et al. Sixteen-week multidrug regimen versus cyclophosphamide, doxorubicin, and fluorouracil as adjuvant therapy for node-positive, receptor-negative breast cancer: an Intergroup study. *J Clin Oncol* 1998; 16:2382–91.
 24. Levine MN, Guyatt GH, Gent M, De Pauw S, Goodyear MD, Hryniuk WM, et al. Quality of life in stage II breast cancer: an instrument for clinical trials. *J Clin Oncol* 1988; 6:1798–810.
 25. Fairclough DL, Fetting JH, Cella D, Wonson W, Moinpour CM. Quality of life and quality adjusted survival for breast cancer patients receiving adjuvant therapy. Eastern Cooperative Oncology Group (ECOG). *Qual Life Res* 1999; 8:723–31.
 26. Fairclough DL, Peterson HF, Cella D, Bonomi P. Comparison of several model-based methods for analysing incomplete quality of life data in cancer clinical trials. *Stat Med* 1998; 17:781–96.

30 Estimation of Cumulative Incidence in the Presence of Competing Risks

*Ted A. Gooley, Wendy Leisenring,
John Crowley, and Barry E. Storer*

CONTENTS

30.1	Introduction	557
30.2	Estimation in the Absence of Competing Risks: Kaplan–Meier Estimate	559
30.3	Estimation in the Presence of Competing Risks: Cumulative Incidence vs. Kaplan–Meier	561
30.4	Example from Real Data: Squamous Cell Carcinoma	563
30.5	Discussion	564
	References	564

30.1 INTRODUCTION

In many fields of medical research, time-to-event end-points are used to assess the potential efficacy of a treatment. The outcome of interest associated with some of these end-points may be a particular type of failure, and it is often of interest to estimate the probability of this failure by a specified time among a particular population of patients. For example, the occurrence of end-stage renal failure is an outcome of interest among patients with insulin-dependent diabetes mellitus (IDDM). Given a sample drawn from the population of patients with IDDM, one may therefore wish to obtain an estimate of the probability of developing end-stage renal failure. As other examples, consider the probability of death due to prostate cancer among patients afflicted with this disease and the probability of the occurrence of CMV retinitis among patients with AIDS.

The examples above share two features. First, the outcome of interest in each is a time-to-event end-point, i.e., one is interested not only in the occurrence of the

outcome but also the time at which the outcome occurs. Secondly, multiple modes of failure exist for each of the populations considered, namely failure from the outcome of interest plus other types whose occurrences preclude the failure of interest from occurring. In the IDDM example, death without end-stage renal failure is a type of failure in addition to the failure of interest, and if a patient with IDDM dies without renal failure, this failure precludes the outcome of interest (end-stage renal failure) from occurring. For the prostate cancer example, deaths from causes unrelated to this cancer comprise types of failure in addition to death from the index cancer, and the occurrence of these alternate failure types preclude the outcome of interest (death from prostate cancer) from occurring. Finally, patients with AIDS who die without CMV retinitis are not capable of going on to develop CMV retinitis, i.e., to fail from the cause of interest. In each of these examples, a competing type of failure exists in addition to the failure of interest. These competing causes of failure are referred to as competing risks for the failure of interest. Specifically, we shall define a competing risk as an event whose occurrence precludes the occurrence of the failure type of interest.

The method due to Kaplan and Meier¹ was developed to estimate the probability of an event for time-to-event end-points, but the assumptions required to make the resulting estimate interpretable as a probability are not met when competing risks are present. Nonetheless, this method is often used and the resulting estimate misinterpreted as representing the probability of failure in the presence of competing risks. Statistical methods for obtaining an estimate that is interpretable in this way are not new,²⁻⁹ and this topic has also received some attention in medical journals.¹⁰⁻¹³ We shall refer to such an estimate as a cumulative incidence (CI) estimate,³ although it has also been referred to as the cause-specific failure probability, the crude incidence curve, and cause-specific risk.¹⁴ Similarly, the term cumulative incidence has been used for various purposes. Our reference to this term, however, will be consistent with its interpretation as an estimate of the probability of failure.

Despite its recognition in both the statistical and medical literature as the appropriate tool to use, CI is not uniformly used in medical research¹⁵ for purposes of estimation in the setting in which competing risks exist. We feel the reason for this is due to both a lack of complete understanding of the Kaplan–Meier method and a lack of understanding of how CI is calculated and hence the resulting difference between the two estimators. In this article, we describe in a nontechnical fashion the Kaplan–Meier estimate not commonly seen. We feel this expression is useful for purposes of understanding why the Kaplan–Meier method results in an estimate that is not interpretable as a probability when used in the presence of competing risks. In addition, this alternate characterization will be extended in a way that allows us to represent CI in a manner similar to that used to obtain the estimate from the Kaplan–Meier method, and in so doing, the validity of CI and the difference between the estimators will be clearly demonstrated.

In the next section, we shall describe the estimate associated with the Kaplan–Meier method in the alternate manner alluded to above in the setting in which competing risks do not exist. The discussion reviews the concept of censoring and provides a heuristic description of the impact that a censored observation has on the estimate. An example based on hypothetical data is used to illustrate the concepts discussed. The subsequent section contains a description of how the two estimates

are calculated when competing risks are present utilizing the description of censoring provided in the preceding section. Data from a clinical trial are then used to calculate each estimate for the end-point of disease progression among patients with head and neck cancer, thus providing further demonstration of the concepts discussed previously. We close with a discussion that summarizes and presents conclusions and recommendations.

30.2 ESTIMATION IN THE ABSENCE OF COMPETING RISKS: KAPLAN–MEIER ESTIMATE

For time-to-event data without competing risks, each patient under study will either fail or survive without failure to last contact. We use *failure* here and throughout as a general term. The specific type of failure depends on the end-point being analyzed. A patient without failure at last contact is said to be censored due to lack of follow-up beyond this time; i.e., it is known that such a patient has not failed by the time of last contact, but failure could occur later.

The most reasonable and natural estimate of the probability of failure by a pre-specified point in time is the simple ratio of the number of failures divided by the total number of patients, provided all patients without failure have follow-up to this time. This simple ratio is appropriately interpreted as an estimate of the probability of failure. This estimate is not only intuitive but is also unbiased when all patients who have not failed have follow-up through the specified time; unbiasedness is a desirable statistical property for estimators.

If one or more patients are censored prior to the specified time, the simple ratio is no longer adequate, and methods that take into account data from the censored patients are required to obtain an estimate consistent with this ratio. The method due to Kaplan and Meier was developed for precisely this purpose, and when competing risks are not present, this method leads to an estimate that is consistent with the desired simple ratio. The resulting estimate is also exactly equal to this ratio when all patients have either failed or have been followed through the specified follow-up time. When used to estimate the probability of failure, one uses the complement of the Kaplan–Meier estimate, which we shall denote by 1-KM, where the Kaplan–Meier estimate (KM) represents the probability of surviving without failure. 1-KM can be interpreted as an estimate of the probability of failure when competing risks are not present.

In order to appreciate how data from patients without follow-up to the specified time are incorporated into 1-KM, it is necessary to understand how censored observations are handled computationally. We present below a heuristic description of censoring not commonly seen. We feel that the use of this approach leads to a clear understanding of what 1-KM represents. In addition, this alternate explanation is used in the following section to explain why 1-KM fails as a valid estimate of the probability of failure and to highlight how and when 1-KM and CI differ when used in the presence of competing risks. What follows is a nontechnical description. The interested reader can find a detailed mathematical description elsewhere.¹⁶

Note that the probability of failure depends on the time-point at which the associated estimate is desired, and as failures occur throughout time, the estimate will

increase with each failure. Because an estimate that is consistent with the simple ratio described above is desired, any estimate that achieves this goal will change if and only if a patient fails. Moreover, if it is assumed that all patients under study are equally likely to fail, it can be shown that each failure contributes a prescribed and equal amount to the estimate, provided all patients have either failed or been followed to a specified time point. This prescribed amount is simply the inverse of the total number of patients under study.

If a patient is censored prior to the time point of interest, however, failure may occur at a time beyond that at which censoring occurred, and this information must be taken into account in order to obtain an estimate consistent with the desired simple ratio. One way to view the manner in which censored observations are handled is as follows. As stated above, each patient under study possesses a potential contribution to the estimate of the probability of failure, and each time a patient fails the estimate is increased by the amount of the contribution of the failed patient. Since patients who are censored due to lack of follow-up through a specified time remain capable of failure by this time, however, the potential contribution of these patients cannot be discounted. In particular, one can consider the potential contribution of a censored patient as being redistributed among all patients known to be at risk of failure beyond the censored time, as noted by Efron.¹⁷ It can be shown that this redistribution to the right makes the resulting estimate consistent with the simple ratio and therefore interpretable as an estimate of the probability of failure. Because of this redistribution, any failure that takes place after a censored observation contributes slightly more to the estimate than do failures prior to the censored observation; i.e., the potential contribution of a patient to the estimate increases after the occurrence of a censored patient. Another way to understand the impact of censoring is to consider such an observation as a *projected* failure or survivor to the time at which the next failure occurs, where this projection is based on the experience of patients who have survived to the time at which censoring occurred.

To help illustrate the above discussion, consider the following hypothetical example summarized in Table 30.1. Suppose that a selected group of 100 patients with IDDM is being followed for development of end-stage renal failure and that none of these 100 dies without failure, i.e., no competing-risk events occur. Assume that all patients have complete follow-up to 15 years after diagnosis of IDDM and 5 of the 100 have end-stage renal failure by this time. The 15-year estimate of renal failure is therefore $5 \times (1/100) = 0.05$, where $1/100$ is the potential contribution of each patient to the estimate of the probability of failure. Suppose the remaining 95 patients survive without renal failure to 20 years, but 25 of these 95 have follow-up only to 20 years; i.e., each of these 25 patients is censored at 20 years. Because each censored patient could develop renal failure beyond 20 years, the potential contribution of each patient cannot be discounted. In particular, the potential contribution to the estimate for each can be thought of as being uniformly redistributed among the 70 patients known to be at risk of failure beyond the censored time. In other words, each remaining patient's potential contribution is increased over $1/100$ by $25 \times (1/100)$ divided among the 70 remaining patients, or $(1/100) + 25 \times (1/100) \times (1/70) = 0.0136$. Suppose 14 of these 70 go on to develop renal failure by 30 years, so that a total of 19 patients are known to have failed by this time. Because 25

TABLE 30.1
Hypothetical Data Illustrating Impact of Censoring on Incidence Estimate

*Event Time	†# Known to Be at Risk	‡# Known Failures (Total # Known Failures by Time	§# Censored	**Contribution of Next Failure	††Incidence Estimate
0	100	0	0	$1/100 = 0.01$	0
15	95	5 (5)	0	0.01	$5(1/100) = 0.05$
20	70	0 (5)	25	$0.01 + 25(0.01)(1/70)$ $= 0.0136$	0.05
30	56	14 (19)	—	—	$0.05 + 14(0.0136)$ $= 0.24$

* Denotes the time at which various events occurred. An event is either a failure or a censored observation.

† Denotes the number of patients who have survived to the associated time, i.e., the number at risk of failure beyond this time.

‡ Denotes the number of patients who failed at the associated time. Parenthetically is the total number of known failures by the associated time.

§ Denotes the number of patients censored at the associated time. Each of these patients could fail at a later time, so the potential contribution to the estimate due to these patients is redistributed evenly among all patients known to be at risk beyond this time.

** Denotes the amount that each failure occurring after the associated time will contribute to the estimate.

†† Gives the estimate of the probability of failure by the associated time. This is derived by multiplying the number of failures at each time by the associated contribution and summing the results or by summing the number of known failures and the number of projected failures and dividing by 100, i.e., the number initially at risk.

patients had follow-up less than 30 years, however, the estimate of failure by this time should be larger than $19/100 = 0.19$; 5 patients fail whose contribution to the estimate is 0.01, and 14 fail whose contribution is 0.0136. The estimate consistent with the desired simple ratio is therefore $5 \times (0.01) + 14 \times (0.0136) = 0.24$. An alternate way to understand this is that because 20% (14/70) of the patients known to be at risk of renal failure beyond 20 years failed by 30 years, it is reasonable to assume that 20% of the 25 patients censored at 20 years will fail by 30 years. This leads to a *projected* total of 24 failures (5 + 14 known failures + 5 *projected* failures) and a 30-year estimate of renal failure of $24/100 = 0.24$.

30.3 ESTIMATION IN THE PRESENCE OF COMPETING RISKS: CUMULATIVE INCIDENCE VS. KAPLAN–MEIER

Consider now the situation in which competing risks exist. In this setting, three outcomes are possible for each patient under study; each will fail from the event of interest, fail from a competing risk, or survive without failure to last contact. In this setting, 1-KM is not capable of appropriately handling failures from a competing risk because in its calculation patients who fail from a competing risk are treated in

the same manner as patients who are censored. Patients who have not failed by last contact retain the potential to fail, however, whereas patients who fail from a competing risk do not. As a result of this, failures from the event of interest that occur after failures from a competing risk contribute more to the estimate than is appropriate in the calculation of 1-KM, as will be demonstrated in the example below. This overestimate is a result of the fact that the potential contribution from a patient who failed from a competing risk, and who is therefore not capable of a later failure, is redistributed among all patients known to be at risk of failure beyond this time. This redistribution has the effect of inflating the estimate above what it should be; i.e., 1-KM is not consistent with the desired simple ratio that is an appropriate estimate of this probability.

An alternative to 1-KM in the presence of competing risks is the cumulative incidence estimate (CI). This estimate is closely related to 1-KM, and patients who are censored due to lack of follow-up are handled exactly as is done in the calculation of 1-KM. Failures from a competing risk, however, are dealt with in a manner appropriate for purposes of obtaining an estimate interpretable as a probability of failure. In the calculation of CI, patients who fail from a competing risk are correctly assumed to be unable to fail from the event of interest beyond the time of the competing-risk failure. The potential contribution to the estimate for such patients is therefore not redistributed among the patients known to be at risk of failure; i.e., failures from a competing risk are not treated as censored as in the calculation of 1-KM. The difference between 1-KM and CI, therefore, comes about from the way in which failures from a competing risk are handled. If there are no failures from a competing risk, 1-KM and CI will be identical. If failures from a competing risk exist, however, 1-KM is always larger than CI at and beyond the time of first failure from the event of interest that follows a competing-risk failure.

Returning to the above example on end-stage renal failure, suppose that competing-risk events do occur, i.e., there are deaths without renal failure. Suppose all assumptions are the same as before with the exception that the 25 patients censored above instead die without renal failure at 20 years. The estimate 1-KM at 30 years is the same as previously because these 25 patients are treated as censored at 20 years. Because competing risks have occurred, however, CI is the appropriate estimate, and the 25 patients who die without renal failure should not be treated as censored. In this simple example, all patients have complete follow-up to 30 years; i.e., each has either developed renal failure, died without failure, or survived without failure by 30 years. The estimate of the probability of renal failure by 30 years should therefore be the number of failures divided by the number of patients, or $19/100 = 0.19$, i.e., the desired simple ratio. Due to the inappropriate censoring of the patients who died without renal failure, however, $1\text{-KM} = 0.24$, an estimate that is not interpretable as the probability of end-stage renal failure.

This simple example illustrates how 1-KM and CI differ when competing risks are present. It also demonstrates why treating patients who fail from a competing risk as censored leads to an estimate (i.e., 1-KM) that cannot be validly interpreted as a probability of failure. In general, the calculation of 1-KM and CI is more involved than shown in the above example due to a more complex combination of event times, but the concepts detailed above are identical.

30.4 EXAMPLE FROM REAL DATA: SQUAMOUS CELL CARCINOMA

To further illustrate the differences between 1-KM and CI, consider the following example taken from a phase III Southwest Oncology Group (SWOG) clinical trial. The objectives of this study were to compare the response rates, treatment failure rates, survival, and pattern of treatment failure between two treatments for patients with advanced-stage resectable squamous cell carcinoma of the head and neck.¹⁸ The two treatments considered were a conventional (surgery and postoperative radiotherapy) and an experimental (induction chemotherapy followed by surgery and postoperative radiotherapy) treatment. We shall use data from this clinical trial among patients treated with the conventional treatment to calculate both 1-KM and CI for the outcome of disease progression.

Among 175 patients entered into the study, 17 were ruled ineligible. Of the 158 eligible patients, 76 were randomized to receive the conventional treatment, and 32 had disease progression while 37 died without progression. Therefore 32 of 76 patients failed from the event of interest (disease progression), while 37 of 76 patients failed from the competing risk of death without progression. The remaining 7 patients were alive without progression at last follow-up and were therefore censored. Each of the seven censored patients had follow-up to at least 7.0 years, and all cases of progression occurred prior to this time. All patients therefore have complete follow-up through 7.0 years after randomization so that the natural estimate of the probability of progression by 7.0 years is $32/76 = 42.1\%$, precisely the value of CI (Figure 30.1). In comparison, the value of 1-KM at this time is 51.6%, as shown in Figure 30.1, the discrepancy being due to the difference in the way that patients who failed from the competing risk of death without progression are handled.

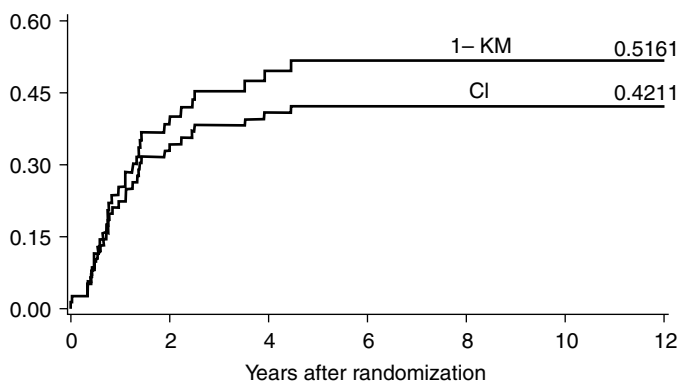


FIGURE 30.1 The complement of the Kaplan–Meier estimate (1-KM) and the cumulative incidence estimate (CI) of disease progression among 76 patients with head and neck cancer. The numerical values of each estimate are indicated.

30.5 DISCUSSION

We have shown that when estimating the probability of failure for end-points that are subject to competing risks, 1-KM and CI can result in different estimates. If it is of interest to estimate the probability of a particular event that is subject to competing risks, 1-KM should never be used, even if the number of competing-risk events is relatively small (and therefore the two estimates not much different.).

It is not clear what 1-KM in such situations represents. The only way to interpret 1-KM in this setting is as the probability of failure from the cause of interest in the hypothetical situation where there are no competing risks and the risk of failure from the cause of interest remains unchanged when competing risks are removed. Because of the way the Kaplan–Meier method handles failures from a competing risk, 1-KM will not be a consistent estimate of the probability of failure from the cause of interest. The discrepancy between 1-KM and CI is dependent on the timing and frequency of the failures from a competing risk; the earlier and more frequent the occurrences of such events, the larger the difference. Regardless of the magnitude of this difference, however, reporting 1-KM in such situations, if it is interpreted as an estimate of the probability of failure, is incorrect and can be very misleading.

The wide availability of statistical software packages that can calculate KM estimates but do not directly calculate CI estimates undoubtedly contributes to the frequent misuse of 1-KM. Although we have not seen the CI estimate offered commercially in any software packages, its calculation is not computationally difficult and programs that accomplish this are reasonably straightforward to write.

The focus of this article was to demonstrate that the methods discussed above lead to different estimates in the presence of competing risks even though each is intended to measure the same quantity. Nonetheless, we feel it crucial to emphasize that presenting only results describing the probability of the event of interest falls short of what should be examined so that one can fully appreciate factors that affect the outcome. When analyzing and presenting data where competing risks occur, it is therefore important to describe probabilities of failure not only from the event of interest but also failures due to competing-risk events. One approach to dealing with this problem is to present an estimate of the time to the minimum of the different types of failure. For a discussion of related topics see Pepe et al.¹⁰

We have focused purely on the estimation of probability of failure in this paper. It is often of interest, however, to compare outcome between two treatment groups. How such comparisons are made and exactly what is compared can be complicated issues and have not been addressed here. Such topics are beyond the scope of this chapter but have been addressed in previous work.^{19,20} If estimation is the goal and competing risks exist, however, the use of 1-KM is inappropriate and yields an estimate that is not interpretable. In such situations, CI should always be used for purposes of estimation.

REFERENCES

1. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1957; 53:457–81.

2. Aalen O. Nonparametric estimation of partial transition probabilities in multiple decrement models. *The Annals of Statistics* 1978; 6:534–45.
3. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. New York: John Wiley; 1980.
4. Prentice RL, Kalbfleisch JD. The analysis of failure times in the presence of competing risks. *Biometrics* 1978; 34:541–54.
5. Benichou J, Gail MH. Estimates of absolute cause-specific risk in cohort studies. *Biometrics* 1990; 46:813–26.
6. Dinse GE, Larson MG. A note on semi-Markov models for partially censored data. *Biometrika* 1986; 73:379–86.
7. Lawless JF. *Statistical Models and Methods for Lifetime Data*. New York: John Wiley; 1982.
8. Gaynor JJ, Feuer EJ, Tan CC, Wu DH, Little CR, Straus DJ, et al. On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. *Journal of the American Statistical Association* 1993; 88:400–9.
9. Pepe MS, Mori M. Kaplan–Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine* 1993; 12:737–51.
10. Pepe MS, Longton G, Pettinger M, Mori M, Fisher LD, Storb R. Summarizing data on survival, relapse, and chronic graft-versus-host disease after bone marrow transplantation: motivation for and description of new methods. *British Journal of Haematology* 1993; 83:602–7.
11. McGiffin DC, Naftel DC, Kirklin JK, Morrow WR, Towbin J, Shaddy R, et al. Predicting outcome after listing for heart transplantation in children: comparison of Kaplan–Meier and parametric competing risk analysis. *Pediatrics* 1997; 16:713–22.
12. Caplan RJ, Pajak TF, Cox JD. Analysis of the probability and risk of cause-specific failure. *International Journal of Radiation, Oncology, Biology, Physics* 1994; 29:1183–6.
13. Gelman R, Gelber R, Henderson C, Coleman CN, Harris JR. Improved methodology for analyzing local and distant recurrence. *Journal of Clinical Oncology* 1990; 8:548–55.
14. Cheng SC, Fine JP, Wei LJ. Prediction of cumulative incidence function under the proportional hazard model. *Biometrics* 1998; 54:219–28.
15. Niland JC, Gebhardt JA, Lee J, Forman SJ. Study design, statistical analyses, and results reporting in the bone marrow transplantation literature. *Biology of Blood and Marrow Transplantation* 1995; 1:47–53.
16. Gooley TA, Leisenring W, Crowley JC, Storer BE. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in Medicine* 1999; 18:695–706.
17. Efron, B. The two sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, University of California Press, 1967; 831–53.
18. Schuller DE, Metch B, Stein DW, Mattox D, McCracken JD. Preoperative chemotherapy in advanced resectable head and neck cancer: final report of the Southwest Oncology Group. *Laryngoscope* 1988; 98:1205–11.
19. Pepe MS. Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association* 1991; 86:770–78.
20. Gray RJ. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *Annals of Statistics* 1988; 16:1141–54.

31 Pitfalls in the Design, Conduct and Analysis of Randomized Clinical Trials

Richard J. Stephens

CONTENTS

31.1	Introduction	568
31.2	Statistics	568
31.3	Key Questions	570
31.3.1	Has the Trial been Designed to Answer a Clear Unconfounded Question?	570
31.3.2	What is the Control Treatment?	570
31.3.3	Are there Predefined Hypotheses for all Key End Points? Are the Number of Statistical Tests Limited, and If not, Have the Significance Levels been Adjusted Accordingly?	571
31.3.4	Do the Eligibility Criteria Cover all the Patients Who are Likely to be Treated this Way Outside of the Trial?	572
31.3.5	Is the Sample Size Based on Information that is Sensible and Feasible?.....	573
31.3.6	Are the Details of the Interim Analyses and Stopping Rules Clearly Laid out?.....	573
31.3.7	Are all Randomized Patients Included and Accounted for in all Analyses?	574
31.3.8	What do the Survival Comparisons Really Show?	575
31.3.9	Have the Hazard Ratio and Especially the 95% Confidence Intervals been Given?	576
31.3.10	Has the Result been put into Context of Previous Work in this Area?.....	577
31.4	Conclusions	577
	References	577

31.1 INTRODUCTION

Randomized clinical trials are generally considered the foundation on which evidence-based medicine is built. In theory they are very simple: half of the patients receive standard treatment; half receive the new treatment; and the two groups are compared in terms of efficacy. But in practice they can actually be very complex. However, just because a trial is randomized does not guarantee its quality, and so the aim of this chapter is to highlight some of the pitfalls that can, and all too often do, exist in the design, conduct, and analysis of randomized trials. In this chapter we will work through a list of 10 key questions that may be a useful way of separating the gold standard randomized trials from the rest.

31.2 STATISTICS

It is first important to explain some of the statistics underpinning randomized trials because without this understanding the implications of many of the pitfalls discussed cannot be fully appreciated.

Classically a randomized trial compares a new experimental therapy with the current standard therapy in an attempt to find out if the new treatment is better and, if so, to estimate how much better. Usually, in cancer the primary end point of interest is survival, but in addition progression-free survival, response, toxicity, palliation, quality of life, and cost-effectiveness may also be important factors in deciding whether the new treatment is better.

If we were able to study every patient with the disease under scrutiny, we would be able to obtain an accurate picture of the disease and treatment. Of course, this is very rarely the case. We usually only have access to a sample of these patients, and all we can do is give an estimation of the true picture. Thus, it is important to remember that whenever we observe a sample of patients they may be drawn from anywhere within the full population, and thus, different studies of patients with the same disease and receiving the same treatment may result in different outcomes. It stands to reason, therefore, that the larger the number of patients we study the better the estimation.

The beauty of randomization is that it ensures that a sample of patients is divided into groups that are as comparable as possible. Given sufficient patients, the groups will not only be automatically matched on obvious characteristics (for example age and sex), but most importantly, in every other aspect. It is the latter point that makes the act of randomization so crucial, and the use of historical controls so risky, as we are still unable to predict with any great accuracy which patients will do well, which badly, and what factors will influence outcome. Randomization thus ensures that the only difference between the groups will be the treatment they receive. Remember, though, that randomization in a trial ensures balance within the chosen subgroup of patients, it cannot ensure balance between the sample chosen and those not being studied.

Given that we are studying a sample of patients, there are a number of statistical terms that describe how close the result from a trial, which is always going to be an estimate, is likely to be to the true result. The power of a trial relates to the chances

of identifying a difference between treatments if it exists. Trials that are under-powered (i.e., do not include enough patients to reliably detect the difference of interest) may therefore result in a false-negative result (also referred to as a type II error). Generally trials are powered at 80–90%, but this still means that 10–20 out of every 100 trials so powered will be a false-negative (i.e., although a difference exists between the groups, the trial suggests no difference). Unfortunately, of course, we never know which negative results are false-negatives.

The p value indicates how likely it is that an observed treatment difference has been found purely by chance; i.e., although no true difference exists between the groups, the trial suggests a difference. Thus, a p value of 0.05 indicates that this result would have occurred by chance 5 times in every 100. It is generally considered that a difference with a p value of ≤ 0.05 is a true and positive result. However it is vital to remember that this actually means that 5 out of every 100 positive results will be false-positives (also referred to as a type I error) found purely by chance. Again, the trouble is that we never know which positive trials are false-positives.

The hazard ratio (HR) is used to describe the difference in the relative risk of death, with, conventionally, a value of <1 indicating a lower risk with the new treatment and >1 indicating a higher risk with the new treatment. Thus, in a survival comparison an HR of 0.85 indicates that the risk of death with the new treatment is 0.85 or, put another way, results in a 15% better survival. Similarly a HR of 1.02 would indicate that with the new treatment survival is 2% worse. A ballpark method of converting the HR into real time is that HR is approximately equal to the median survival of patients on the standard treatment divided by the median survival of patients on the new treatment. In addition the HR is approximately equal to the natural log of the proportion of patients surviving at a particular time-point on the standard treatment, divided by the natural log of the proportion of patients surviving at the same time-point on the new treatment. Thus, for example, if the median and 1 year survival of patients on a standard treatment are 9 months and 20%, respectively, and the HR from a trial is 0.85, the estimated median and 1 year survival for patients on the new treatment are approximately 10.6 months and 25.5% respectively. This assumes the survival curves both follow an exponential shape; more complex methods are needed to describe situations where curves cross (e.g., in trials of surgery versus no surgery, the surgery group may experience a worse postoperative period but a better long term outcome).

However, when describing survival differences between treatments, probably the most important statistical term is the 95% confidence interval (CI) of the HR. Given that we know that the result of our trial can only be an estimate of the true result, the 95% CI indicates the range in which we are 95% sure that the true value lies. Thus, an HR of 0.85 with 95% CI of 0.65 and 1.05 indicates that we are 95% confident that the new treatment is somewhere between 35% better and 5% worse. This surprisingly wide range is the sort of range commonly obtained from randomized trials with a sample size of about 250 patients. Thousands of patients are required to obtain confidence intervals with a range of about 10%. Even in a trial of over 1000 patients, comparing surgery with or without adjuvant chemotherapy in patients with non-small cell lung cancer (NSCLC), Scagliotti et al.¹ reported a HR of 0.96 with 95% CI 0.81–1.31, indicating that, compared to the median survival of 48 months with

surgery alone, in their trial the best estimate they could make was that the true effect on survival of adding adjuvant chemotherapy lay somewhere between a detriment of 5.5 months and a benefit of 11 months.

31.3 KEY QUESTIONS

With this background understanding of the statistics, we can now consider some of the key pitfalls that can occur in a randomized trial, although sometimes *pitfalls* is the wrong word as trials can of course be deliberately designed or analyses deliberately performed to weigh the scales in favor of one treatment or another. Nevertheless the aim of this chapter is to alert readers to the major deficiencies that can occur in trial design and trial reporting and that may prevent the trial from being a true and unbiased comparison of the treatments.

31.3.1 HAS THE TRIAL BEEN DESIGNED TO ANSWER A CLEAR UNCONFOUNDED QUESTION?

Trials should always aim to answer only one clear question. In an ideal world, a clean trial design would be to take a standard chemotherapy regimen and either add or replace one drug in the experimental treatment. It will then be clear that any benefit or detriment observed in the trial will be directly related to the drug in question. In reality, of course, such clean designs are not always possible or desirable. Often two widely used treatments need to be compared or entirely different schedules or combinations of treatments investigated. Nevertheless, results from such trials that, for instance, change more than one drug (or schedules or doses) often leave the question unanswered as to the relative value of each changed factor. For example Kelly et al.² in their trial in patients with advanced NSCLC compared paclitaxel and carboplatin (PC) given in 3-weekly cycles and vinorelbine and cisplatin (VC) given in 4-weekly cycles. The trial reported equal efficacy, but if say, a benefit had been seen with PC would this have been as a result of the use of paclitaxel, carboplatin, or the 3-weekly cycles? When trials are clearly designed to answer one question, there is often a temptation to explore the data further, but it is always important not to over-interpret the result. Thus if in a pragmatic trial a choice of chemotherapy regimen is permitted within one arm, the relative benefits of the chemotherapy regimens used cannot and should not be untangled.

31.3.2 WHAT IS THE CONTROL TREATMENT?

What should guide trial design is equipoise, or the uncertainty principle, which perhaps might be judged by the willingness of clinicians to be enrolled themselves should they develop that condition. In a randomized trial the choice of control treatment is paramount. Logically it should always be the current best standard treatment for the condition, although often knowing what is acknowledged as best is difficult. Nevertheless, it is not difficult to see that the choice of the control treatment can significantly influence the way the trial result is interpreted, as unfortunately much more attention is paid to trials with a positive result. Thus, in order to increase the chances of attaining a positive outcome, trials can be designed to compare the new

treatment with a poor or inappropriate control. A common trick is to compare the new treatment alone with the new treatment in combination with a standard treatment. Thus, in lung cancer there are examples of trials comparing a new drug versus that new drug plus cisplatin; for example, Splinter et al.³ compared teniposide with or without cisplatin in advanced NSCLC. Cisplatin is a very effective drug, and not surprisingly, the trial showed that the cisplatin combination resulted in improved response, progression-free survival, and overall survival. Such results can be used to claim that the new combination should be considered as an effective standard treatment, irrespective of whether the new drug actually has any useful effect or not.

While most randomized trials are designed to compare a new treatment against a standard treatment, trials may also be designed to assess whether a new treatment is equivalent to a standard treatment (for example the new treatment may have preferable attributes, such as being given orally rather than intravenously or being less costly) or seeing which of two standard treatments is better. Because of the difficulty, due to the huge numbers of patients required, of showing that a new treatment is equivalent (i.e., the 95% CI need to be very tight around the HR) to a standard treatment, a course of action sometimes taken is to show that the new treatment is better than a previous standard to the same degree as the current standard. Thus if treatment B is 5% better than treatment A, the options for designing a trial to look at new treatment C are either to try to show that C is equivalent to B or that C is also 5% better than treatment A. If in a trial of C versus A, treatment C turns out to be 5% better (or perhaps 2% or 8% better), this leaves the key question of what should be considered the gold standard unanswered and creates difficulties for future trials because the control treatment to be used is unclear. For example in a meta-analysis⁴ reported in 1995, a highly statistically significant survival benefit was seen in patients with advanced NSCLC when cisplatin-based chemotherapy was added to supportive care ($p < 0.0001$). Therefore when Roszkowski et al.⁵ started their trial later in 1995 to look at the use of docetaxel in patients with advanced NSCLC, should they have used supportive care alone or supportive care plus cisplatin-based chemotherapy as their control regimen? They chose the former, but this now leaves unanswered the question of the relative benefits of docetaxel compared to cisplatin-based chemotherapy.

The fact that the control arm should be a widely accepted standard treatment also raises interesting issues about placebo-controlled trials. The placebo effect may be something that needs to be controlled for in a trial where the primary end point is subjective (e.g., pain control), but in cancer trials with a survival outcome, using a placebo is not standard clinical practice and is unlikely to affect survival, so is it necessary or, indeed, ethical?

31.3.3 ARE THERE PREDEFINED HYPOTHESES FOR ALL KEY END POINTS? ARE THE NUMBER OF STATISTICAL TESTS LIMITED, AND IF NOT, HAVE THE SIGNIFICANCE LEVELS BEEN ADJUSTED ACCORDINGLY?

While we need to be aware that a proportion of positive trial results may in fact be false-positives, this problem also affects analyses within a trial, as the more tests that are performed the more likely it is that they will be contaminated with false results. To reduce this risk a clear number of end points (and related analyses) should be

predefined in the protocol to limit the number of statistical tests performed in a trial, and thus to limit the chance of false-positive results. A good way of doing this is to consider that within a trial there is only a certain amount of p value spending. So, if one test is performed and the result is $p \leq 0.05$, then the result can be considered significant. If two tests are performed, then perhaps they should only be considered statistically significant if $p \leq 0.025$, or as is often used to accommodate interim analyses, the first is only considered significant if $p \leq 0.001$ so that the second can be considered significant if $p \leq 0.049$. It is important when reading publications to be aware of how many statistical tests have been performed and what is claimed to be significant, as there can often be a tendency to trawl the data for interesting results and significant p values. Often such difficulties particularly affect the assessment of quality of life (QOL) where patients may complete long questionnaires at a number of time-points. Consider, for example, one table relating to the assessment of QOL in the paper by Sundstrom et al.⁶ where 84 p values are calculated, although the authors recognized the problem and indicate that only $p < 0.01$ would be considered statistically significant.

It is, of course, logical to list the pretreatment characteristics and to highlight balance (or imbalance) between groups. However, it is illogical to apply statistical tests to show balance or imbalance. Statistical tests are used to estimate the likelihood that an observed difference has not occurred by chance. However, differences in pretreatment characteristics can only have occurred by chance, and it is thus an inappropriate use of a statistical test and a wasteful use of p value spending. Recent examples of this unnecessary testing can be found in papers by Tada et al.⁷ and Langendijk et al.⁸ If imbalances in pretreatment characteristics and especially in stratification factors are observed, the analysis of the key end points should probably be adjusted accordingly (see [Chapter 28](#)).

All subgroup analyses should be viewed with caution as results may be affected not only by multiple p value testing but also by reduced sample size. Only predefined analyses where the hypothesis has been clearly stated upfront and sample sizes calculated should be regarded as reliable.

31.3.4 DO THE ELIGIBILITY CRITERIA COVER ALL THE PATIENTS WHO ARE LIKELY TO BE TREATED THIS WAY OUTSIDE OF THE TRIAL?

The results of trials will influence the way future patients are treated. It is therefore important that the eligibility criteria reflect this population of patients because it is unlikely that all the eligibility criteria will be remembered and adhered to once the trial is over. Unrepresentative samples are commonly seen in phase II studies where a very select group of patients is chosen. These may often be the younger, fitter patients who are most likely to be able to cope with increased treatment complexity or side effects of a new treatment. As a result, impressive phase II results are often not reproducible in randomized phase III trials. It is important that in the continued quest for improved survival we do not concentrate on smaller and smaller subgroups of patients and overlook the vast majority. For example, the median age of patients in large multicenter randomized trials of relatively straightforward chemotherapy for patients with advanced colorectal cancer is often around 65 years, whereas the median age of patients presenting with this disease is usually 75 years.

Reflecting the population to be treated post-trial extends to the exclusion of patients in a trial post-randomization. Patients are often excluded from analyses as a result of trial-specific investigations (e.g., reference pathology) that would not normally be performed once the trial is completed. Therefore, the trial result on a “pure” population may not reflect the result attainable in normal clinical practice.

31.3.5 IS THE SAMPLE SIZE BASED ON INFORMATION THAT IS SENSIBLE AND FEASIBLE?

All too often sample sizes are based on what is feasible rather than what is realistic. For instance, we know that in lung cancer the addition of a new modality, be it radiotherapy or chemotherapy, will probably only improve survival by about 5%. For instance, the NSCLC meta-analysis⁴ showed that the addition of cisplatin-based chemotherapy to surgery increased 5-year survival by 5%, and the meta-analysis exploring the benefit of adding radiotherapy to chemotherapy in small cell lung cancer⁹ also reported an improvement in 3-year survival of 5%. Therefore, it is unrealistic to consider that as a result of tinkering with the drugs, dosages, or schedules, we are suddenly going to see advantages of a further 10 or 15%. Yet the vast majority of lung cancer trials are based on seeing differences of about 15%, which will generally require around 400 patients. Some even aim for larger effects. For example, Ranson et al.¹⁰ powered their trial in advanced NSCLC to look for an improvement from 20% survival at 1 year with supportive care to 40% with paclitaxel, and Sculier et al.¹¹ considered a 75% relative increase in median survival might be possible with accelerated 2-weekly chemotherapy compared to standard 3-weekly chemotherapy for patients with extensive SCLC. A question then arises as to whether it is ethical to run any trial of less than perhaps 1000 patients given the high probability of an inconclusive — or worse, a false-positive or false-negative — result. An even greater dilemma occurs with equivalence trials. Taking the same example that the addition of a modality improves survival by about 5%, what happens when we want to show that a new treatment is as effective as standard? If we compare the new therapy to standard therapy with a trial of 400 patients, we may finish with a HR of around 1.00 but with 95% CI of about $\pm 15\%$. So all we could conclude is that the new treatment is somewhere between 15% better and 15% worse than standard and thus could actually be 10% worse than the old therapy. For example, Kelly et al.² claimed that in their trial PC was equally efficacious to VC even though a 12% survival detriment or a 25% benefit with PC could not be ruled out.

31.3.6 ARE THE DETAILS OF THE INTERIM ANALYSES AND STOPPING RULES CLEARLY LAID OUT?

To ensure patient safety, it is imperative that the accumulating data are reviewed at regular intervals throughout the trial. Whether *regular* means annually, when accrual reaches certain targets, or when certain numbers of events have occurred will depend on the design of the trial. The interim data should be reviewed by a committee of clinicians and statisticians, who usually will be completely independent of, and not involved with, any other aspect of the trial. However arguments have been put forward that such committees should not be completely independent.¹² Rules for when the trial

should close early must also be agreed on, and there are a number of options, from fixed p values to Bayesian statements such as “the evidence must convince sceptics”. It is important that among such data monitoring and ethics committee (DMEC) members there is knowledge of the disease and treatments and previous DMEC experience, as often DMECs will be called upon to make very difficult decisions. There are numerous examples where trials have stopped early, but the results have been unconvincing and required new trials to be set up to clarify the situation. For example, two trials of neoadjuvant chemotherapy for NSCLC^{13,14} both stopped early after accruing 60 patients because the neoadjuvant arm was showing a highly statistically significant benefit ($p < 0.008$ and $p < 0.001$ for the two trials, respectively). However, subsequently several large trials have had to be set up to clarify whether any benefit exists.

Few trial reports list stopping rules, but guidelines on stopping suggest p values of at least < 0.001 need to be seen at interim analyses to avoid unnecessary closure. Certainly trials that stop because a p value of < 0.05 was observed at an interim analysis need to be questioned.

31.3.7 ARE ALL RANDOMIZED PATIENTS INCLUDED AND ACCOUNTED FOR IN ALL ANALYSES?

A good policy is to account for every patient in every analysis. Thus including categories such as *not assessed* or *died* in tables and reporting the numbers of patients (not just the proportions) makes all analyses completely transparent to the reader.

The easiest and most logical group to analyze is everyone who has been randomized. This is the strict definition of intent to treat. At the time of randomization all patients should have been considered suitable for the treatments being studied and thus will reflect the population who are likely to be offered the treatment post-trial. Papers often list subgroups of patients who are excluded from analyses: for example, post-randomization investigations or independent review may show patients to be ineligible; patients do not receive any or all of their protocol treatment; or patients are not assessed for an end point. However, removing patients for any of the above reasons has the potential to bias the analysis sample. For example, in trials in advanced NSCLC, Georgoulas et al¹⁵ excluded 35 of the 441 patients randomized and all analyses (that were claimed to be intention-to-treat) were then performed on the remaining 406 patients, and Schiller et al.¹⁶ considered 52 of 1207 patients to be ineligible post-randomization. An alternative argument is that removing ineligible patients ensures that the results of the trial better reflect the population defined by the eligibility criteria. Either way, usually the proportion excluded is small enough not to make significant differences to the result.

Accounting for all patients (whether it be the intention to treat or eligible population) extends to all analyses including, of course, response and toxicity. It is important to report the response rate as the proportion of patients who achieve complete or partial response out of the total number of patients in the group. Quoting the response rate as just the proportion of patients who have been assessed at a certain time-point may mask the fact that patients may have had to stop the treatment due to toxicity or death.

Similarly the most logical and widely used method to report toxicity is to quote the proportion of patients with grade 3 or 4 for each key symptom within a defined

time period from randomization. Such an analysis will inevitably include some noise because patients will have had symptoms pretreatment and some patients will have toxicity as a result of non-protocol treatment. However, understanding and applying the concept of intent to treat is important, as the trial should be trying to record the experiences of the whole group of patients chosen to receive a certain treatment. If some patients do not actually receive the protocol treatment and have to receive different treatment, perhaps with different side-effects, that is a key message. In virtually all analyses it is much better to report the proportion of patients with a good or bad experience rather than the mean or median score. The mean or median can mask or dilute the fact that a small proportion of patients had good or bad experiences.

31.3.8 WHAT DO THE SURVIVAL COMPARISONS REALLY SHOW?

It is very easy to quickly observe survival plots and make assumptions about the between-treatment differences. However before jumping to conclusions it is important to consider the detail. Survival should always include all patients randomized, be calculated from the date of randomization, include all causes of death, be measured by constructing Kaplan–Meier curves, and be compared by using the logrank test. In addition, plots should indicate the number of patients at risk over time, and overall survival should be reported using the HR and 95% CI. Taking the start date as anything but randomization (which is the one common time-point for all patients) will have the potential to bias the result. For example, the date of diagnosis may not be accurate for all patients, or the date of start of treatment may include different delays for different groups and does not exist for patients who do not start treatment.

Although the cause of death may be of interest to the trialists as an indication of how the treatment is working, in a real sense this may be much less important to the patient. Thus, survival analyses that only report deaths from cancer may be interesting but very misleading. For example, in a cancer-specific survival analysis, a treatment that causes many early treatment-related deaths may appear to be the better treatment. Thus, the analysis of disease-specific survival rates in the paper by Sundstrom et al.⁶ and of progression-free survival in which patients who died from causes unrelated to disease or treatment were censored¹⁷ may be difficult to interpret.

Many papers purport to show differences between treatments in terms of time to progression with the use of a Kaplan–Meier plot, taking progression as the event and censoring those alive (or dead) without progression. This sort of analysis can be very misleading as patients who fail from a competing risk (for example an early treatment-related death) that precludes the possibility of achieving the event are treated the same as surviving patients who still have the potential to progress (see [Chapter 28](#)). Recent examples of this can be found in papers by Sunstrom et al.,⁶ Ranson et al.¹⁰, and Pujol et al.¹⁸ Progression-free survival, which takes into account deaths without progression, should always be the preferred analysis.

It is often interesting to explore whether any overall survival difference observed is consistent across all subgroups. Analyses stratified for pretreatment characteristics are therefore useful. While Sause et al.¹⁹ did just that, their subgroup analyses do not appear to have been predefined or accounted for in the sample size and so should only be considered as exploratory or hypothesis generating. While exploratory

analyses are acceptable, analysis by post-randomization factors (such as treatment received or response) in general are not, as the groups being compared may be defined by the outcome being tested. Thus, for example, comparing the survival of responders versus non-responders is flawed because the responders have to survive long enough to respond. Therefore, analyses such as those presented by Fukuoka et al.²⁰ comparing survival by responders and Socinski et al.²¹ showing survival by number of cycles of chemotherapy received must be viewed with great caution. If such comparisons are required, then one method is to take all the patients alive at a landmark time-point (say, 6 months from randomization) and analyze from this new baseline, thus moving the event of interest (response, number of cycles of chemotherapy, etc.) to pre-baseline.

31.3.9 HAVE THE HAZARD RATIO AND ESPECIALLY THE 95% CONFIDENCE INTERVALS BEEN GIVEN?

Reporting survival at a landmark time-point, be it median or 1 year survival, rather than the HR may be misleading. For example, in a trial of surgery versus a non-surgical intervention, the expectation may well be that the surgery group is likely to result in high early postoperative mortality but better longer-term survival. Thus reporting survival at, say, 1 month or at 5 years might give an inaccurate picture of the true between-treatment difference. Similarly some survival curves may overlap for a considerable time before splitting; see for example Fossella et al.²² and Takada et al.²³

In most situations the HR and 95% CI will be the best estimate of the overall survival difference between any two treatments (although as discussed in the statistics section above, an overall HR may not be appropriate if the curves clearly cross). Unfortunately, all too often results are presented only as *p* values with no indication of the HR, and often the HR is quoted without the 95% CI. The true result is not the HR but rather that we are 95% sure that the true result is somewhere in the range of the 95% CI. Table 31.1 shows the results of a number of recent trials comparing surgery with surgery plus adjuvant chemotherapy in patients with advanced NSCLC. Simply considering the *p* values, which range from 0.006 to 0.9, leads one to the

TABLE 31.1

Surgery vs. Surgery + Adjuvant Platinum-Based Chemotherapy. Summary Statistics from the NSCLC Meta-Analysis and Subsequent Trials that Included >250 Patients

Trial/MA (ref)	N	<i>p</i>	HR	95% CI
NSCLC MA (4)	1394	0.08	0.87	0.74–1.02
IALT (24)	1867	<0.03	0.86	0.76–0.98
ALPI (1)	1088	0.59	0.96	0.81–1.13
BLT (25)	381	0.90	1.02	0.77–1.35
JBR10 (26)	459	0.006	0.85	0.81–0.92
CALGB 9633 (27)	344	0.03	0.62	0.41–0.95

conclusion that some of these trials are positive and some negative, and there has been a lot of discussion as to the cause of these perceived differences. Even the HRs look very different, ranging from 0.62 to 1.02, but when the 95% CI are considered, it is clear that they all stride HR values of 0.81–0.92, and therefore the results could all be considered completely consistent with one another.

31.3.10 HAS THE RESULT BEEN PUT INTO CONTEXT OF PREVIOUS WORK IN THIS AREA?

Trials are rarely islands. Results need to be presented and discussed in the context of the totality of previous work. However, Clarke and Chalmers²⁸ reviewed the discussion sections of reports of trials published in five major journals during one month in 1997 and found that only 2 of 26 placed their results in the context of an up-to-date systematic review. Repeating this exercise in 2001,²⁹ they reported no improvement, with only 3 of 30 trials being so reported. Such findings are disappointing and suggest that there is a general lack of awareness that individual trials are only part of the whole picture. We must never lose sight of the fact that cancer is a global problem and without global collaboration progress will continue to be painfully slow.

31.4 CONCLUSIONS

All trials and all trial results are important because they all in some way advance the progress of human knowledge. However, trials are far from equal. It is therefore important to be aware of the numerous pitfalls that can, and do, appear in the design, conduct, and analysis of randomized trials, so that the reliability of the evidence they provide can be judged. Hopefully the 10 questions explored in this chapter will give the reader some insight into the pitfalls that can affect the reliability of trial results. At the end of the day our ultimate aim as trialists is to improve the treatment of future patients, and it is important therefore that the whole research community is as rigorous and honest in our work as we can be.

REFERENCES

1. Scagliotti, G.V., Fossati, R., Torri, V., et al. Randomized study of adjuvant chemotherapy for completely resected stage I, II, or IIIa non-small cell lung cancer. *J Natl Cancer Inst*, 95, 1453, 2003.
2. Kelly, K., Crowley, J., Bunn, P.A. Jr., et al. Randomized phase III trial of paclitaxel plus carboplatin versus vinorelbine plus cisplatin in the treatment of patients with advanced non-small cell lung cancer: a Southwest Oncology Group trial. *J Clin Oncol*, 19, 3210, 2001.
3. Splinter, T.A., Sakhmoud, T., Festen, J., et al. Two schedules of teniposide with or without cisplatin in advanced non-small cell lung cancer: a randomized study of the European Organization for Research and Treatment of Cancer Lung Cancer Cooperative Group. *J Clin Oncol*, 14, 127, 1996.
4. Non-Small Cell Lung Cancer Collaborative Group. Chemotherapy in non-small cell lung cancer: a meta-analysis using updated data on individual patients from 52 randomised clinical trials. *BMJ*, 311, 899, 1995.

5. Roszkowski, K., Pluzanska, A., Krzakowski, M., et al. A multicenter, randomized, phase III study of docetaxel plus best supportive care versus best supportive care in chemotherapy-naïve patients with metastatic or non-resectable localized non-small cell lung cancer (NSCLC). *Lung Cancer*, 27, 145, 2000.
6. Sundstrom, S., Bremnes, R.M., Kaasa, S., et al. Cisplatin and etoposide regimen is superior to cyclophosphamide, epirubicin and vincristine regimen in small cell lung cancer: results from a randomized phase III trial with 5 years' follow-up. *J Clin Oncol*, 20, 4665, 2002.
7. Tada, H., Tsuchiya, R., Ichinose, Y., et al. A randomized trial comparing adjuvant chemotherapy versus surgery alone for completely resected pN2 non-small cell lung cancer (JCOG9304). *Lung Cancer*, 43, 167, 2004.
8. Langendijk, H., de Jong, J., Tjwa, M., et al. External irradiation versus external irradiation plus endobronchial brachytherapy in inoperable non-small cell lung cancer: a prospective randomized study. *Rad Oncol*, 58, 257, 2001.
9. Pignon, J.P., Arriagada, R., Ihde, D.C., et al. A meta-analysis of thoracic radiotherapy for small cell lung cancer. *N Engl J Med*, 327, 1618, 1992.
10. Ranson, M., Davidson, N., Nicolson, M., et al. Randomized trial of paclitaxel plus supportive care versus supportive care for patients with advanced non-small cell lung cancer. *J Natl Cancer Inst*, 92, 1074, 2000.
11. Sculier, J.P., Paesmans, M., Lecomte, J., et al. A three-arm phase III randomised trial assessing, in patients with extensive disease small cell lung cancer, accelerated chemotherapy with support of haematological growth factor or oral antibiotics. *Br J Cancer*, 85, 1444, 2001.
12. Harrington, D., Cowley, J., George, S.L., et al. The case against independent monitoring committees. *Stat Med*, 13, 1411, 1994.
13. Roth, J.A., Fossella, F., Komaki, R., et al. A randomized trial comparing perioperative chemotherapy and surgery with surgery alone in resectable stage IIIa non-small cell lung cancer. *J Natl Cancer Inst*, 86, 673, 1994.
14. Rosell, R., Gomez-Condina, J., Camps, C., et al. A randomized trial comparing pre-operative chemotherapy plus surgery with surgery alone in patients with non-small cell lung cancer. *N Engl J Med*, 330, 153, 1994.
15. Georgoulas, V., Papadakis, E., Alexopoulos, A., et al. Platinum-based and non-platinum-based chemotherapy in advanced non-small cell lung cancer: a randomised multicentre trial. *Lancet*, 357, 1478, 2001.
16. Schiller, J.H., Harrington, D., Belani, C.P., et al. Comparison of four chemotherapy regimens for advanced non-small cell lung cancer. *N Engl J Med*, 346, 92, 2002.
17. Shepherd, F.A., Giaccone, G., Seymour, L., et al. Prospective, randomized, double-blind, placebo-controlled trial of marimastat after response to first-line chemotherapy in patients with small cell lung cancer: a trial of the national Cancer Institute of Canada — Clinical Trials Group and the European Organization for Research and Treatment of Cancer. *J Clin Oncol*, 20, 4434, 2002.
18. Pujol, J.-L., Daures, J.-P., Riviere, A., et al. Etoposide plus cisplatin with or without the combination of 4'-epidoxorubicin plus cyclophosphamide in treatment of extensive small cell lung cancer: a French Federation of Cancer Institutes multicenter phase III randomized study. *J Natl Cancer Inst*, 93, 300, 2001.
19. Sause, W., Kolesar, P., Taylor, S. IV, et al. Final results of phase III trial in regionally advanced unresectable non-small cell lung cancer. *Chest*, 117, 358, 2000.
20. Fukuoka, M., Yano, S., Giaccone, G., et al. Multi-institutional randomized phase II trial of gefitinib for previously treated patients with advanced non-small cell lung cancer. *J Clin Oncol*, 21, 2237, 2003.

21. Socinski, M.A., Schell, M.J., Peterman, A., et al. Phase III trial comparing a defined duration of therapy versus continuous therapy followed by second-line therapy in advanced stage IIIB/IV non-small cell lung cancer. *J Clin Oncol*, 20, 1335, 2002.
22. Fossella, F.V., DeVore, R., Kerr, R.N., et al. Randomized phase III trial of docetaxel versus vinorelbine or ifosfamide in patients with advanced non-small cell lung cancer previously treated with platinum-containing chemotherapy regimens. *J Clin Oncol*, 18, 2354, 2000.
23. Takada, M., Fukuoka, M., Kawahara, M., et al. Phase III study of concurrent versus sequential thoracic radiotherapy in combination with cisplatin and etoposide for limited stage small cell lung cancer: results of the Japan Clinical Oncology Group Study 9104. *J Clin Oncol*, 20, 3054, 2002.
24. The International Adjuvant Lung Cancer Trial Collaborative Group. Cisplatin-based adjuvant chemotherapy in patients with completely resected non-small cell lung cancer. *N Engl J Med*, 350, 351, 2004.
25. Waller, D., Peake, M.D., Stephens, R.J., et al. Chemotherapy for patients with non-small cell lung cancer: the surgical setting of the Big Lung Trial. *Eur J Cardiothor Surg*, 26, 173, 2004.
26. Winton, T.L., Livingstone, R., Johnson, D., et al. A prospective randomised trial of adjuvant vinorelbine (VIN) and cisplatin (CIS) in completely resected stage IB and II non-small cell lung cancer (NSCLC). Intergroup JBR.10. *Proc Am Soc Clin Onc*, Late Breaking Abstracts Booklet, 17, abs. 7018, 2004.
27. Strauss, G.M., Herndon, J., Maddaus, M.A., et al. Randomised clinical trial of adjuvant chemotherapy with paclitaxel and carboplatin following resection in stage IB non-small cell lung cancer (NSCLC): Report of Cancer and Leukemia Group B (CALGB) protocol 9633. *Proc Am Soc Clin Onc*, Late Breaking Abstracts Booklet, 17, abs. 7019, 2004.
28. Clarke, M., and Chalmers, I. Discussion sections in reports of controlled trials published in general medical journals: islands in search of continents? *JAMA*, 280, 280, 1998.
29. Clarke, M., Alderson, P., and Chalmers, I. Discussion sections in reports of controlled trial published in general medical journals. *JAMA*, 287, 2799, 2002.

32 Dose-Intensity Analysis

Joe Pater

CONTENTS

32.1	Introduction	581
32.2	Relating Delivered Dose to Outcome in Individual Patients on a Trial	582
32.2.1	Problems in Analyzing Results in Patients Who Stopped Treatment Before Its Intended Conclusion	582
32.2.2	Confounding	583
32.3	Analyses Combining Data from Multiple Studies	584
32.4	Designing Trials to Test the Dose Intensity Hypothesis	585
32.4.1	Settings for Testing	586
32.4.2	Design of Individual Trials	586
	References	587

32.1 INTRODUCTION

Despite the inherent methodologic difficulties to be discussed in this chapter, analyses that have attempted to relate the intensity of cytotoxic chemotherapy to its effectiveness have had substantial influence both on the interpretation of data from cancer clinical trials and on recommendations for practice in oncology. Prior to a discussion of analytic pitfalls, a brief description of the evolution of the concept of dose intensity and its impact on the design and analysis of trials will be presented. It should be noted from the outset that although most discussions of these issues involve the term *dose intensity*, this term has been used in many ways. In fact, precisely how the amount of treatment given over time is quantified is an important problem in itself.

The publication that first provoked interest in the issue of the importance of delivering adequate doses of chemotherapy was a secondary analysis by Bonnadonn¹ of his randomized trials of cyclophosphamide, methotrexate and 5-Fluorouracil (CMF) in the adjuvant therapy of women with breast cancer. This analysis subdivided patients in the trials according to how much of their protocol prescribed chemotherapy they actually received. There was a clear positive relationship between this proportion and survival. Patients who received 85% or more of their planned protocol dose had a 77% 5 year relapse free survival compared to 48% in those who received less than 65%. The authors concluded, “it is necessary to administer combination chemotherapy at a full dose to achieve clinical benefit.”

Shortly afterward, interest in the role of dose intensity was further heightened by a series of publications by Hryniuk and his colleagues. Instead of examining

outcomes of individual patients on a single trial, Hryniuk's approach was to correlate outcomes of groups of patients on different trials with the planned dose intensity of their treatment. In this case, the intensity of treatment was related not to the protocol specified dose but to a standard regimen. In trials both in breast cancer and ovarian cancer, the results of these analyses indicated a strong positive correlation between dose intensity and outcome. Thus, in the treatment of advanced breast cancer the correlation between planned dose intensity and response was 0.74 ($p < 0.001$) and between actually delivered dose intensity and response 0.82 ($p < 0.01$).² In the adjuvant setting a correlation of 0.81 ($p < 0.00001$) was found between dose intensity and 3 year relapse free survival.³ In ovarian cancer the correlation between dose intensity and response was 0.57 ($p < 0.01$).⁴

Reaction to these publications was quite varied. While some authors pointed out alternate explanations for the results, most appeared to regard these analyses as having provided insight into an extremely important concept. For example, the then director of the National Cancer Institute, Vincent DeVita, in an editorial commentary on Hryniuk and Levine's paper on adjuvant therapy, concluded "a strong case can be made for the point of view that the most toxic maneuver a physician can make when administering chemotherapy is to arbitrarily and unnecessarily reduce the dose".⁵ Hryniuk himself argued, "since dose intensity will most likely prove to be a major determinant of treatment outcomes in cancer chemotherapy, careful attention should be paid to this factor when protocols are designed or implemented even in routine clinical practice".⁶

Despite numerous trials designed to explore the impact of dose intensity on outcomes of therapy over the more than 20 years since Bonnadonna's publication, the topic remains controversial. Thus, a recent book entitled *Ovarian Cancer: Controversies in the Management*⁷ devotes a chapter to a debate on the topic. Similar discussions continue within other sites.

The remainder of this chapter will review the methodologic and statistical issues involved in the three settings mentioned above; namely, studies aimed at relating the delivery of treatment to outcomes in individual patients, studies attempting to find a relationship between chemotherapy dose and outcome among a group of trials, and finally, trials aimed at demonstrating in a prospective randomized fashion an impact of dose or dose intensity.

32.2 RELATING DELIVERED DOSE TO OUTCOME IN INDIVIDUAL PATIENTS ON A TRIAL

Three years after the publication of Bonnadonna's paper, Redmond, Fisher, and Wieand⁸ published a methodologic and statistical critique of the analysis and interpretation of his findings. In my view, subsequent discussions of this issue have been mainly amplifications or reiterations of Redmond's article, so it will be summarized in detail here.

32.2.1 PROBLEMS IN ANALYZING RESULTS IN PATIENTS WHO STOPPED TREATMENT BEFORE ITS INTENDED CONCLUSION

Cancer chemotherapy regimens are usually given over at least several months, so no matter the setting, some patients are likely to develop disease recurrence or progression

prior to the planned completion of therapy. How to calculate the expected amount of treatment in such patients is a problem. If the expected duration is considered to be that planned at the outset, patients who progress on treatment will by definition receive less than full doses and a positive relation between completeness of treatment and outcome will be likely.

An obvious solution to this problem is to consider as expected only the duration of treatment up to the time of recurrence, which was the approach taken by Bonnadonna. However, this method leads to a potential bias in the opposite direction because, generally speaking, toxicity from chemotherapy is cumulative and doses tend to be reduced over time. Thus, patients who stop treatment early may have received a higher fraction of their expected dose over that time period. In fact, in applying this method to data from an NSABP trial, Redmond et al. found an inverse relationship between dose and outcome; i.e., patients who received higher doses over the time they were on treatment were *more* likely to develop disease recurrence.

A third approach suggested by Redmond et al. was to use what has become known as the landmark⁹ method. That is, the effect of delivered dose up to a given point in time is assessed only among patients who were followed and were free of recurrence until that time. This method avoids the biases mentioned above but omits from the analysis patients who recur during the time of greatest risk of treatment failure. The application of this method to the same NSABP trial indicated no effect of variation in the amount of drug received up to 2 years on subsequent disease free survival.

Finally, Redmond et al. proposed using dose administered up to a point in time as a time dependent covariate recalculated at the time of each recurrence in a Cox proportional hazards model. In their application of this technique to NSABP data, however, they found a significant delivered dose effect in the placebo arm, a finding reminiscent of the well-known results of the clofibrate trial.¹⁰ The apparent explanation for this result is that treatment may be delayed or omitted in the weeks preceding the diagnosis of recurrence. Thus, when drug delivery in the two months before diagnosis of recurrence was not included in the calculation, the dose effect disappeared both in the placebo and active treatment arms.

32.2.2 CONFOUNDING

The methodologic problems of analyses that attempt to relate completeness of drug administration to outcome are not confined to the difficulties of dealing with patients who progress while on therapy. Citing the clofibrate trial mentioned above, Redmond et al. also pointed out that even if one could calculate delivered vs. expected dose in an unbiased manner, there would still remain the problem that patients who comply with or who tolerate treatment may well differ from those who do not on a host of factors that themselves might relate to ultimate outcome. Thus, the relationship between administered dose and outcome might be real in the sense that it is not a product of bias in the way the data are assembled or analyzed but might not be causal as it is the product of confounding by underlying patient characteristics. Since the clinical application of an association of dose delivery and outcome rests on the assumption that the relationship is causal, the inability to draw a causal conclusion is a major concern.¹¹

Analyses similar to those of Bonnadonna have been repeated many times,¹² sometimes with positive and sometimes with negative results. However, it is hard to argue with the conclusion of Redmond et al.'s initial publication; i.e., this issue will not be resolved by such analyses. Only a prospective examination of the question can produce definitive findings.

32.3 ANALYSES COMBINING DATA FROM MULTIPLE STUDIES

As mentioned earlier, instead of relating outcomes of individual patients to the amount of treatment they received, Hryniuk et al. assessed whether the outcomes of groups of patients treated with regimens that themselves varied in intended or delivered dose intensity correlated with that measure. In order to compare a variety of regimens used, for example, for the treatment of metastatic breast cancer, Hryniuk calculated the degree to which the rate of drug administration in mg/M²/week corresponded to a standard protocol and expressed the average of all drugs in a given regimen as a single proportion, relative dose intensity. (See Table 32.1 for an example of such a calculation). He then correlated this quantity calculated for individual arms of clinical trials or for single arm studies with a measure of outcome, such as response or median survival.

Hryniuk's work has been criticized on two primary grounds, first, the method of calculating dose intensity, and second, the appropriateness of his method of combining data. With respect to the method of calculation it has been pointed out that the choice of a standard regimen is arbitrary and different results are obtained if different standards are used.^{12,13} Furthermore, the approach ignores the impact of drug schedule and makes untested assumptions about the therapeutic equivalence of different drugs. In more recent work,¹⁴ Hryniuk accepted these criticisms and addressed them by calculating a new index called summation dose intensity. The intensity of each agent is incorporated into this index by considering what proportion of the dose necessary to produce a 30% response rate is being used. All drugs in a regimen are credited in this fashion. This index is more empirically based and avoids reference to an arbitrary standard. As Hryniuk points out, whether it will resolve the debate about dose intensity will depend upon its prospective validation in randomized trials.

Hryniuk's work has also been considered a form of meta-analysis and criticized because of its failure to meet standards¹⁵ for such studies. In my view, with the exception of an article by Meyer, Hryniuk, and Goodyear,¹⁶ these studies are not actually meta-analyses. Meta-analyses are conventionally considered to involve

TABLE 32.1
Calculation of Dose Intensity

Drug	Dose	mg/M ² /wk	Reference mg/M ² /wk	Dose Intensity
Cyclophosphamide	100 mg/M ² days 1–14	350	560	0.62
Methotrexate	40 mg/M ² days 1 and 8	20	28	0.71
Fluorouracil	600–600 mg/M ² days 1 and 8	300	480	0.62
Dose intensity for regimen				0.65

combining data from a set of studies, each of which attempted to measure some common parameter, for example, an effect size or odds ratio. Hryniuk's work, on the other hand, compares results from one study to the next, and estimates a parameter — the correlation between dose intensity and outcome — that is not measured in any one study. Irrespective of this distinction, however, the criticism that Hryniuk's early studies did not clearly state how trials were identified and selected seems valid.

Formal meta-analytic techniques have, however, been applied to this issue. As mentioned, the article by Meyer et al. contains what I would consider a true meta-analysis of trials assessing the role of dose intensity in patients with non-Hodgkin's lymphoma. An accompanying editorial pointed out,¹³ however, that the arms of these trials differed in more than dose intensity. Thus, because its conclusions were otherwise based on non-randomized comparisons, the editorial argued that even in this analysis no (in Sackett's¹⁷ terminology) level I evidence had been developed to support the role of dose intensity.¹³ In fact, despite the fact that they drew on data from randomized trials, Hryniuk's studies generated only what Sackett considers level V evidence because they relied mostly on comparisons involving nonconcurrent, nonconsecutive case series.

A much more extensive attempt to generate more conclusive evidence by combining data from randomized trials testing dose intensity was carried out by Torri, Korn, and Simon.¹⁸ Using standard search techniques, these authors identified all published randomized trials from 1975 to 1989 dealing with the chemotherapy of advanced ovarian cancer. For each arm of each trial they calculated dose with an equalized standard method similar to the summation standard method described above. They used a log linear model that included a separate fixed effect term for each study to estimate the relationship between total dose intensity and the log odds of response and log median survival. The inclusion of the fixed effect terms ensured that comparisons were between the arms of the same trial, not across trials. They also used a multiple linear regression model to assess the relative impact of the intensity of various drugs. Their analysis showed a statistically significant relationship between dose intensity and both outcomes, although the magnitude of the relationship was less than that observed in Hryniuk's studies — a finding they attributed to the bias inherent in across trial comparisons. They concluded, “the validity of the dose intensity hypothesis in advanced ovarian cancer is substantiated based on the utilisation of improved methodology for analysis of available data. This approach suggests hypotheses for the intensification of therapy and reinforces the importance of formally evaluating dose intense regimens in prospective randomised clinical trials.”

32.4 DESIGNING TRIALS TO TEST THE DOSE INTENSITY HYPOTHESIS

Authors critical of Bonnadonna's⁸ or Hryniuk's¹³ methodology called for prospective (randomized) tests of the underlying hypothesis, as did Hryniuk³ himself. However, how difficult this clinical research task actually is does not seem to have been widely appreciated.¹⁹ There are two fundamental problems: 1) the multitude of circumstances in which differences in dose intensity need to be tested, and 2) the difficulty of designing individual trials.

32.4.1 SETTINGS FOR TESTING

In order for the hypothesis that dose intensity of chemotherapy is an important determinant of outcome of cancer treatment to be tested in clinical trials, there would need to be consensus on what constitutes convincing evidence for or against the hypothesis. Clearly a single negative trial would not be sufficient because it could always be argued that the clinical setting or the regimen used was not appropriate. Thus, in order to build a strong case against this hypothesis, a large number of trials of different regimens in different settings would have to be completed. Conversely, and perhaps less obviously, a positive trial, while it would demonstrate that the hypothesis is true in a specific circumstance, would not necessarily indicate that maximizing dose intensity in other situations would achieve the same result. As Siu and Tannock put it, “generalizations about chemotherapy dose intensification confuse the reality that the impact of this intervention is likely to be disease, stage, and drug specific”.²⁰ It can be argued, in fact, that systematically attempting to test hypotheses of this type in clinical trials is unlikely to be productive, at least from the point of view of providing a basis for choosing treatment for patients.²¹ In any case, given the number of malignancies for which chemotherapy is used and the number of regimens in use in each of those diseases, the clinical research effort required to explore fully the implications of this hypothesis is enormous. It is perhaps not surprising that the issue is still unresolved.^{14,20}

32.4.2 DESIGN OF INDIVIDUAL TRIALS

Equally problematic is designing trials that clearly delineate the impact of dose intensity on outcome. Dose intensity, as defined by Hryniuk et al., is a rate, the numerator of which is the total dose given and the denominator the time over which it is delivered. Because of the mathematical relationship among the components of this rate (total dose, treatment duration, and dose per treatment), a change in dose intensity can be achieved in different ways and may or may not result in a change in total dose given or in the overall duration of treatment, both of which may themselves have an impact on outcome. Bigonzoli and Piccart have described five possible models for increasing dose intensity by increasing the dose per treatment course or decreasing the interval between courses: 1) increased dose per course with no change in the total dose; 2) increased dose per course and higher cumulative dose; 3) increased dose per course but reduced cumulative dose; 4) reduced dose per course, decreased interval between courses, and constant cumulative dose; 5) decreased interval between courses but constant dose per course and constant cumulative dose. All but one of the models result in the same or higher cumulative dose. Likewise all but one result in a shorter total duration of therapy. Regimens that achieve higher dose intensity by decreasing the interval between treatments without reducing dose per treatment (model 5) are considered *dose dense*, relative to the standard regimens from which they are derived. Many such regimens have been developed and tested as the availability of hematopoietic growth factors made them possible. In one randomized trial,²² increasing the dose density of a standard adjuvant chemotherapy regimen for breast cancer has been demonstrated to improve disease free and overall survival. This contrasts with other approaches to increasing dose intensity, where it has been much easier to show that lowering doses below conventional levels has an adverse effect on outcome than to demonstrate an advantage to increasing

dose above conventional levels.²⁰ Whether this means that dose densification will prove to be the best approach to increasing dose intensity is a question that would, like demonstrating a role for dose intensity itself, require many empiric comparisons to resolve. In my view,¹⁹ considering the difficulties outlined above, Siu and Tannock²⁰ are correct that “further research (in this area) should concentrate on those disease sites and/or chemotherapeutic agents for which there is a reasonable expectation of benefit.” Although the retrospective methods described above are inadequate for drawing causal conclusions, they still may be useful in pointing out where further trials will be most fruitful.

REFERENCES

1. Bonnadonna G, Valagussa B. Dose-response effect of adjuvant chemotherapy in breast cancer. *N Engl J Med* 1981;304:10–15.
2. Hryniuk W, Bush H. The importance of dose intensity in chemotherapy of metastatic breast cancer. *J Clin Oncol* 1984;2:1281–1288.
3. Hryniuk W, Levine MN. Analysis of dose intensity for adjuvant chemotherapy trials in stage II breast cancer. *J Clin Oncol* 1986;4:1162–1170.
4. Levin L, Hryniuk WM. Dose intensity analysis of chemotherapy regimens in ovarian carcinoma. *J Clin Oncol* 1987;5:756–767.
5. Devita VT. Dose-response is alive and well. *J Clin Oncol* 1986;4:1157–1159.
6. Hryniuk WM, Figueredo A, Goodyear M. Applications of dose intensity to problems in chemotherapy of breast and colorectal cancer. *Semin Oncol* 1987;4(suppl 4):3–11.
7. Gershenson DM, McGuire WP. *Ovarian Cancer: Controversies in Management*. Churchill Livingstone. New York 1998.
8. 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.
9. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. *J Clin Oncol* 1983;1:710–719.
10. 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.
11. Geller NL, Hakes TB, Petrone GR, Currie V, Kaufman R. Association of disease-free survival and percent of ideal dose in adjuvant breast chemotherapy. *Cancer* 1990;66:1678–1684.
12. Henderson IC, Hayes DF, Gelman R. Dose-response in the treatment of breast cancer: a critical review. *J Clin Oncol* 1988;6:1501–1513.
13. Gelman R, Neuberg D. Making cocktails versus making soup. *J Clin Oncol* 1991;9:200–203.
14. Hryniuk W, Frei E, Wright FA. A single scale for comparing dose-intensity of all chemotherapy regimens in breast cancer: summation dose-intensity. *J Clin Oncol* 1998;16:3137–3147.
15. L’Abbe KA, Detsky AS, O’Rourke K. Meta-analysis in clinical research. *Ann Intern Med* 1987;107:224.
16. Meyer RM, Hryniuk WM, Goodyear MDE. The role of dose intensity in determining outcome in intermediate-grade non-Hodgkin’s lymphoma. *J Clin Oncol* 1991;9:339–347.

17. Sackett DL. Rules of evidence and clinical recommendations in the use of antithrombotic agents. *Chest* 1989;95(suppl):2S–4S.
18. Torri V, Korn EL, Simon R. Dose intensity analysis in advanced ovarian cancer patients. *Br J Cancer* 1993;67:190–197.
19. Pater JL. Introduction: implications of dose intensity for cancer clinical trials. *Sem Oncol* 1987;14(Suppl 4):1–2.
20. Siu LL, Tannock IP. Chemotherapy dose escalation: case unproven. *J Clin Oncol* 1997;15:2765–2768.
21. Pater JL, Eisenhauer E, Shelley W, Willan A. Testing hypotheses in clinical trials: experience of the National Cancer Institute of Canada Clinical Trials Group. *Cancer Treat Rep* 1986;70:1133–1136.
22. Citron ML, Berry D, Cirincione C, Hudis C, Winer E, Gradishar W, et al. Randomized trial of dose-dense versus conventionally scheduled and sequential versus concurrent combination chemotherapy as postoperative adjuvant treatment of node-positive primary breast cancer: first report of Intergroup Trial C9741/Cancer and Leukemia Group B Trial 9741. *J Clin Oncol* 2003;21:1431–1439.

33 Sequential Randomization

Joe Pater and John Crowley

CONTENTS

33.1	Introduction	589
33.2	Examples	591
33.2.1	CALGB 7461	591
33.2.1.1	Design	591
33.2.1.2	Results.....	591
33.2.1.3	Comment.....	591
33.2.2	SWOG 8624.....	592
33.2.2.1	Design	592
33.2.2.2	Results.....	592
33.2.2.3	Comments	592
33.2.3	ECOG 4494	593
33.2.3.1	Design	593
33.2.3.2	Results.....	593
33.2.3.3	Comments	593
33.2.4	NCIC CTG MY7	593
33.2.4.1	Design	593
33.2.4.2	Results.....	593
33.2.4.3	Comments	594
33.3	Discussion and Recommendations	594
	References	596

33.1 INTRODUCTION

It is a common practice in some cancer sites, particularly in hematologic malignancies, to study two aspects of disease management by randomizing patients on a trial twice, first between two or more initial or induction regimens and then later randomizing those who do or do not reach a particular end point to different maintenance or salvage treatments. This approach differs from a true factorial design because not all patients are randomized to both study questions and, more importantly, because the probability of being included in the second comparison depends upon the outcome of the first. Thus there are all the difficulties of possible treatment interactions associated with factorial designs in cancer clinical trials (see Green, in this volume)¹ as well as the additional pitfall of a potential confounding of the results of the first randomization by those of the second.

The fundamental problem with sequential randomization was described some time ago in a review of clinical trials in acquired immunodeficiency.² The authors pointed out that such patients might often be enrolled on more than one trial in sequence and that this could confuse the interpretation of the initial trial. For example, if one of the therapies on the second trial proved extremely effective and more patients on one of the arms of the first reached an end point that made them eligible for the second, the apparent effectiveness of that arm might actually be due to the treatment studied in the second trial.

As pointed out by Green et al.,³ it is important to distinguish between studies where the goal is to determine which planned sequence of treatments is the best from those where separate early and late therapeutic interventions are of primary interest. This can be an issue even for two-arm trials. For example, in a trial of one vs. two years of adjuvant therapy for breast cancer,⁴ the decision was made to ask the treatment policy question rather than the question of whether to continue an additional year of therapy in those patients who survived one year. This meant randomizing to the two arms up front rather than randomizing only those patients who survived one year of treatment.

Similarly, with separate induction and maintenance treatments, the decision needs to be made as to whether the primary question is which sequence of treatments is the best or whether separate induction and maintenance questions are being asked. In the former situation, randomization up front to the entire sequence of treatments is preferred, but low participation in the subsequent treatments due to noncompliance or planned restriction to a subset of patients can be an issue. For example, in a trial of two chemotherapy options followed by either radiation or chemotherapy for multiple myeloma,⁵ 614 patients were randomized up front, but only 187 were treated in the maintenance phase according to the prerandomized arm. As a practical matter radiation versus chemotherapy was analyzed according to treatment received as essentially a nonrandomized comparison.

Wahed and Tsiatis⁶ have explored the statistical implications of the sequential design from the perspective of the information it provides on which of the potential treatment sequences is superior, and they have developed estimators of survival functions in these circumstances. Their approach assumes, however, that the goal of a sequentially randomized trial is to compare four (or more) specific treatment strategies rather than to address two separate questions, the answers to which might be applied in different circumstances.

If separate induction and maintenance questions are of interest, then a choice of a short term end point for the first question can minimize any potential confounding. In acute leukemia it is widely accepted that a complete response (CR) at induction confers patient benefit, so that a design that randomizes patients at induction to different strategies and then randomizes only patients in CR to maintenance, with a survival or relapse-free survival end point, can answer both questions. The validity of response as a short term end point is less clear in other tumor types, however.^{7,8}

Despite the frequent use of the sequential randomization design, there have been very few published discussions of its properties and the problems it raises for interpretation. In the remainder of this chapter we discuss examples of trials that illustrate these problems and close with some recommendations.

33.2 EXAMPLES

The following examples were chosen to illustrate how sequentially randomized designs have been used in practice.

33.2.1 CALGB 7461

33.2.1.1 Design

This trial⁹ of chemotherapy in patients with multiple myeloma randomized patients among four induction regimens (intravenous or oral melphalan or the nitrosoureas, CCNU or BCNU). Patients whose disease had not progressed and who were still on study were randomized to receive vincristine and prednisone vs. no additional therapy. Five hundred eighty-nine patients were randomized among the four induction regimens. Three hundred two patients were subsequently randomized between vincristine and prednisone and no additional therapy.

33.2.1.2 Results

Survival of patients randomized to vincristine and prednisone measured from the time of the second randomization was significantly prolonged compared to those receiving no additional therapy. Further analyses indicated that this effect was confined to patients receiving melphalan (though no formal test for interaction was described). The authors then combined the induction arms into two categories (melphalan or nitrosourea) and presented separate survival curves (measured from the time of the second randomization) for patients treated with the four possible combinations. The patients receiving melphalan plus vincristine and prednisone had distinctly superior survival compared to those who received a nitrosourea with or without vincristine and prednisone or those who received melphalan without vincristine and prednisone. A time dependent Cox model was used to compare survival on the two melphalan regimens from the time of initial randomization and the combined melphalan regimens to the two nitrosourea arms. The time dependent covariate reflected the treatment of some patients with vincristine and prednisone at the second randomization. No significant differences were found for the induction comparisons adjusted for maintenance therapy.

33.2.1.3 Comment

This study illustrates several issues: 1) Although it appears to be generally assumed that the major problem with the sequential design is that the second randomization confounds the interpretation of the first,¹⁰ in this case it appears there was an interaction between the induction regimens and the second randomization. 2) An analysis was done that compared survival on four regimens that included both randomizations suggesting the authors were, in fact, as assumed by Wahed and Tsiatis, interested in selecting a single superior treatment approach. However, this analysis included only the patients who completed the second randomization. 3) An analysis from the time of initial randomization was done that adjusted for the second randomization, but this

did not provide a basis for an intent-to-treatment assessment of which combination of treatments was superior.

33.2.2 SWOG 8624

33.2.2.1 Design

This trial¹¹ compared three induction chemotherapy regimens (VMCP/VBAP vs. VAD vs. VMCP/VBAPP). Patients who had a 75% regression of their disease on any induction arm were randomized to maintenance interferon or no interferon. Patients who had a lesser response or were stable at 12 months received interferon and dexamethasone. Patients who had no response at six months or who progressed received no additional study therapy. Five hundred nine eligible patients were randomized among the three induction regimens, and 281 eligible patients were registered on the maintenance phase. Of the latter, 193 were randomized between interferon maintenance and no maintenance, and 88 were registered on the single-arm study of interferon/dexamethasone.

33.2.2.2 Results

Overall survival from the time of randomization was superior for the VAD and VMCP/VBAPP arms compared to VMCP/VBAP. This result was attributed to the higher steroid doses in these arms. There was no difference in survival as measured from the time of maintenance registration for patients randomized to interferon or no interferon. Interestingly the majority of patients registered on the interferon/dexamethasone arm came from the least active induction arm — VMCP/VBAP. Overall survival on this arm, assessed from the time of maintenance registration, was superior to that of the patients randomized between interferon and no interferon. The authors point out that this could be due either to the effectiveness of the regimen or to the fact that patients with a slow response to initial therapy have an intrinsically better long term prognosis.

33.2.2.3 Comments

1. The induction and maintenance components of this trial were clearly intended to be separate questions because no analysis was conducted that examined the multiple possible treatment strategies.
2. As the authors point out, the definitions of response in myeloma may not be adequate to support designing trials that use these criteria as a basis for enrolling patients on subsequent study questions (see also reference 7).
3. Although the authors do not make the point, this trial could well be an illustration of confounding by registration on a second study in that the inferiority of the VMCP/VBAP arm might have been even greater if a disproportionate number of patients on this induction treatment had not received what appeared to be an active maintenance regimen. (In fact subsequent studies have indicated that the use of dexamethasone in maintenance is effective).¹²

33.2.3 ECOG 4494

33.2.3.1 Design

This trial¹³ assessed the use of the anti-CD20 antibody rituximab in the treatment of diffuse large cell lymphoma. Patients were randomized to receive standard chemotherapy with or without the addition of rituximab. Responding patients were then randomized to receive maintenance rituximab or no further therapy. Six hundred thirty-two patients were enrolled on the initial comparison, and 415 on the maintenance question.

33.2.3.2 Results

This study has only been published in abstract form. Initial results indicate that the addition of rituximab to induction chemotherapy prolonged time to treatment failure but not overall survival. Likewise maintenance rituximab prolonged time to treatment failure but not survival. However, there was strong evidence for an interaction between the induction and maintenance treatments in that maintenance rituximab was highly, but only, effective in patients who did not receive induction rituximab. The results of the comparison of induction regimens among patients who did and did not receive rituximab maintenance were not presented but can be inferred from the statement that the induction and maintenance questions interact. The authors point out that the study was not designed to compare directly the two induction regimens.

33.2.3.3 Comments

This is an extreme example of how the components of a sequentially randomized trial can confound each other. As in the case of CALGB 7461, it appears that the effect of the maintenance treatment was dependent upon the induction arm — a negative interaction that Girling et al.¹⁰ consider unlikely. Further, the inclusion of a second randomization made an unambiguous interpretation of the induction question impossible.

33.2.4 NCIC CTG MY7

33.2.4.1 Design

This trial^{12,14} addressed the role of high dose pulsed dexamethasone in induction and maintenance of multiple myeloma. Patients were randomized to one of four arms: melphalan prednisone followed by maintenance dexamethasone or no maintenance, or melphalan dexamethasone with the same maintenance alternatives. Five hundred eighty-four eligible patients were randomized on the four arms; 292 did not progress and received their assigned maintenance treatment.

33.2.4.2 Results

The analysis of the induction question comparing the two dexamethasone containing arms to the two prednisone arms showed no difference in survival and greater toxicity on the dexamethasone arms. Randomization to these two arms was halted,

and enrollment on the two prednisone arms was continued until sufficient patients had been included to meet the sample size requirements for the maintenance question. The final comparison of dexamethasone maintenance to no maintenance included only the events occurring among the patients who reached the maintenance phase. This as-treated analysis found a highly significant difference in progression free survival but no difference in overall survival.

33.2.4.3 Comments

This trial addressed a typical induction/maintenance question, but because of the issues raised in this chapter, all patients were randomized to both questions at the outset. In the end, though, it was analyzed in exactly the same way as trials in which the second randomization is delayed. The design had the advantage that if questions of confounding had arisen an intent-to-treat comparison of the four treatment sequences could have been conducted – though with limited power. Although not included in the publications cited, this analysis was done and confirmed the published results.

33.3 DISCUSSION AND RECOMMENDATIONS

As the examples cited above illustrate, practice is not uniform in the design and analysis of sequentially randomized trials. In our view, the following considerations should be taken into account before undertaking such trials or allowing randomization on more than one study:

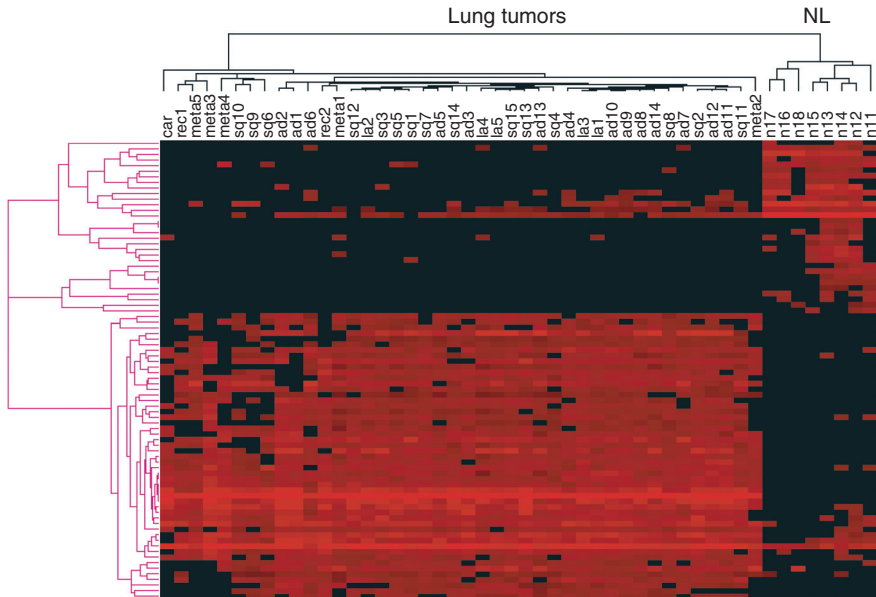
1. As in the case of truly factorial designs, the most important design issue to be resolved prior to initiating a sequentially randomized trial is whether the study questions are independent or not. The problem of independence has to be considered in two ways.
 - a. The first is the usual need for absence of potential interaction in the statistical sense. In two of the trials described above, interaction was observed, and in one case (ECOG 4494) its presence almost totally subverted the trial's interpretation. A disadvantage of the sequentially randomized trial is that, because not all patients are included, the power to test for interaction is limited.
 - b. The second is more contextual and relates to the question of whether sequentially designed trials can be viewed, as Wahed and Tsiatis treated them, as tests of four or more treatment strategies. In some situations this is clearly the case, but in others, as in the example discussed by Girling et al., the investigators see the second randomization as addressing a question that applies outside the context of the initial randomization. The importance of this distinction is its implication for the eventual interpretation of the study. If, in fact, the goal is to compare and select the best group of treatment sequences that are likely to be used in practice, the question of whether a superior result is due to the

first or second component of the sequence becomes somewhat moot. On the one hand, however, one could argue that in such cases patients should be allocated to treatment sequences at the outset, so that an intent-to-treat analysis can be done. On the other hand, if the second question is important in its own right, independently of the regimens tested in the first, randomization obviously should take place at the point where the results would be applied in practice, stratifying for initial treatment. Generally, in this setting the results of the second study question can be interpreted unambiguously. ECOG 4494 might seem to be a counterexample, but the two questions studied in the trial were neither conceptually nor, as it turned out, statistically independent.

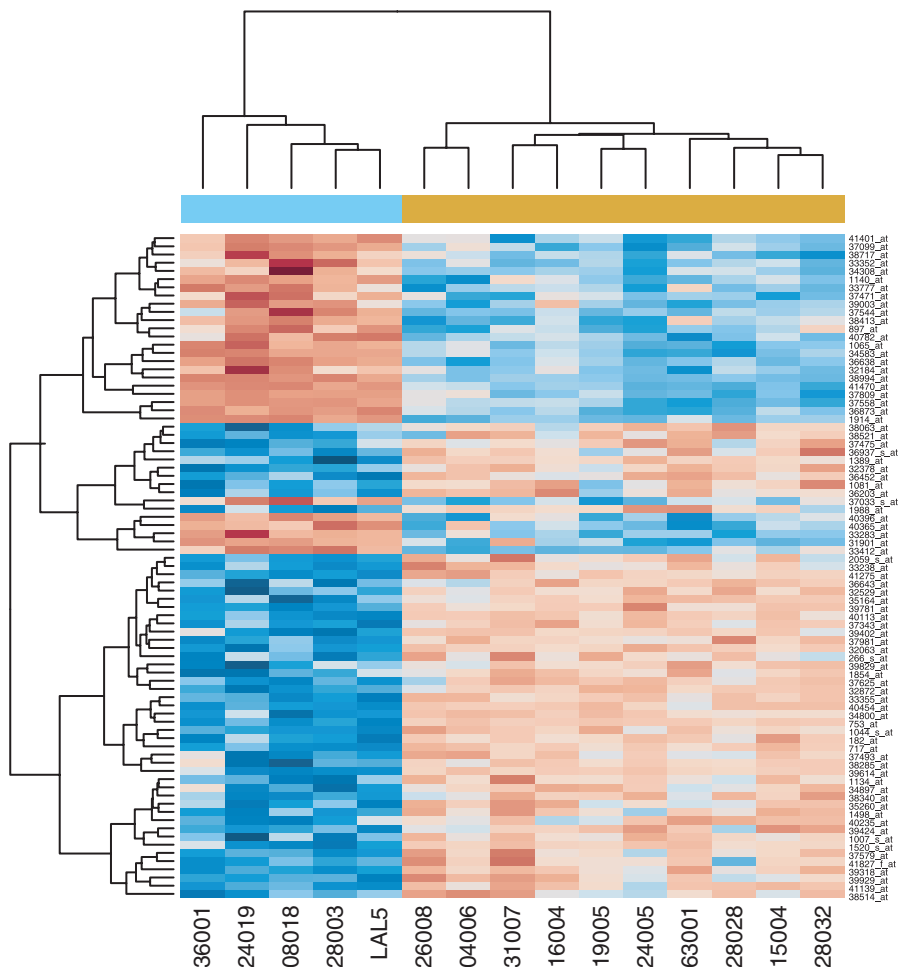
2. It is tempting to believe that purely statistical methods can unambiguously answer questions that trials were not well designed to answer. Wahed and Tsiatis derive methods for analyzing a sequentially randomized study as if the intent of the trial were to compare treatment policies, but if that were the case, randomization is better done up front. The methods proposed are not particularly intuitive, and the optimality properties provided apply only when none of the survival times are censored. Similarly the use of time-dependent covariates to make comparisons between induction treatments from the first randomization, adjusted for maintenance treatments, has many model assumptions for validity and certainly does not enjoy the kind of straightforward interpretation that a randomized trial without a maintenance randomization would.
3. Although these designs clearly have pitfalls and can produce results that are difficult to interpret, it is hard to see how they can be avoided entirely because they address issues that arise in practice. The best patient management may well require the selection of an optimal sequence of therapy. When that is the case, as Girling et al. argue, it is almost certainly preferable to address such issues in a controlled fashion in the context of a planned study than to let practitioner preferences confound results without acknowledgement. On the other hand, a secondary sequentially randomized question of lesser importance should not be added to a planned trial on the assumption that additional information will be obtained at no cost because it is possible that the opportunity to answer the first question unambiguously will be lost. Interactions do happen in practice, and confounding is a real possibility. Sometimes asking one question leads to a clear answer, while asking two questions really leads to no answer. Similar considerations apply to allowing substantial proportions of patients on one study to be enrolled on another, for example, in studies of adjuvant therapy in breast cancer where large studies addressing separate issues such as local management, hormone therapy, and chemotherapy are often enrolling patients in the same geographic area or from the same centers. Clearly, the same problems could arise in this setting if, for example, a substantial number of patients on one arm of an adjuvant chemotherapy trial were selectively enrolled on a comparison of hormone therapies.

REFERENCES

1. Green S.G., Factorial designs with time to event endpoints. This volume.
2. Byar D.P. et al., Design considerations for AIDS trials. *New Engl J Med* 323:1343, 1990.
3. Green S.G., Benedetti J., and Crowley J., *Clinical Trials in Oncology*, 2nd ed., Chapman & Hall/CRC Press, New York, 2003.
4. Salmon S.E. et al., Chemotherapy is superior to sequential hemibody irradiation for remission consolidation in multiple myeloma: a Southwest Oncology Group study. *J Clin Oncol* 8:1575, 1990.
5. Wahed A.S. and Tsiatis A.A., Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics* 60:124, 2004.
6. Tangen, C., Phase II designs with time to event endpoints. This volume.
7. Durie B.G.M. et al., Magnitude of response with myeloma frontline therapy does not predict outcome: importance of time to progression in Southwest Oncology Group chemotherapy trials. *J Clin Oncol* 22:1857, 2004.
8. Cooper M.R. et al. Single, sequential, and multiple alkylating agent therapy for multiple myeloma: a CALGB study. *J Clin Oncol* 4:1331, 1986.
9. Girling D. et al., *Clinical Trials in Cancer*. Chapter 4.6. Oxford University Press, New York, 2003.
10. Salmon S.E. et al., Combination chemotherapy, glucocorticoids, and interferon alpha in the treatment of multiple myeloma: a Southwest Oncology Group study. *J Clin Oncol* 12:2405, 1994.
11. Shustik C. et al., Melphalan-dexamethasone is not superior to melphalan-prednisone as induction therapy in multiple myeloma. *Proc Am Soc Clin Oncol* 20:298a, 2001.
12. Habermann T.M. et al., Phase III trial of rituximab-CHOP (R-CHOP) vs. CHOP with a second randomization to maintenance rituximab (MR) or observation in patients 60 years of age or older with diffuse large B-cell lymphoma. *Blood* 102:8a, 2003.
13. Shustik C. et al., Dexamethasone maintenance versus observation in patients with previously untreated multiple myeloma: a National Cancer Institute of Canada Clinical Trials Group study: MY.7. *Proc Am Soc Clin Oncol* 23:558, 2004.



COLOR FIGURE 25.2 Results from the agglomerative hierarchical clustering algorithm.



COLOR FIGURE 27.4 A heatmap comparing leukemia patients. The patients are the columns; the genes they are clustered on are the rows. Shading indicates relative gene expression.