Recent Advances in Clinical Trial Design and Analysis

Cancer Treatment and Research

Kamps WA, Humphrey GB, Poppema S (eds): Hodgkin's Disease in Children: Controversies and Current Practice. 1988. ISBN 0-89838-372-2

Muggia FM (ed): Cancer Chemotherapy: Concepts, Clinical Investigations and Therapeutic Advances. 1988. ISBN 0-89838-381-1

Nathanson L (ed): Malignant Melanoma: Biology, Diagnosis, and Therapy. 1988. ISBN 0-89838-384-6

Pinedo HM, Verweij J (eds): Treatment of Soft Tissue Sarcomas. 1989. ISBN 0-89838-391-9

Hansen HH (ed): Basic and Clinical Concepts of Lung Cancer. 1989. ISBN 0-7923-0153-6

Lepor H, Ratliff TL (eds): Urologic Oncology. 1989. ISBN 0-7923-0161-7

Benz C, Liu E (eds): Oncogenes. 1989. ISBN 0-7923-0237-0

Ozols RF (ed): Drug Resistance in Cancer Therapy. 1989. ISBN 0-7923-0244-3

Surwit EA, Alberts DS (eds): Endometrial Cancer. 1989. ISBN 0-7923-0286-9

Champlin R (ed): Bone Marrow Transplantation. 1990. ISBN 0-7923-0612-0

Goldenberg D (ed): Cancer Imaging with Radiolabeled Antibodies. 1990. ISBN 0-7923-0631-7

Jacobs C (ed): Carcinomas of the Head and Neck. 1990. ISBN 0-7923-0668-6

Lippman ME, Dickson R (eds): Regulatory Mechanisms in Breast Cancer: Advances in Cellular and Molecular Biology of Breast Cancer. 1990. ISBN 0-7923-0868-9

Nathanson L (ed): Malignant Melanoma: Genetics, Growth Factors, Metastases, and Antigens. 1991. ISBN 0-7923-0895-6

Sugarbaker PH (ed): Management of Gastric Cancer. 1991. ISBN 0-7923-1102-7

Pinedo HM, Verweij J, Suit HD (eds): Soft Tissue Sarcomas: New Developments in the Multidisciplinary Approach to Treatment. 1991. ISBN 0-7923-1139-6

Ozols RF (ed): Molecular and Clinical Advances in Anticancer Drug Resistance. 1991. ISBN 0-7923-1212-0

Muggia FM (ed): New Drugs, Concepts and Results in Cancer Chemotherapy. 1991. ISBN 0-7923-1253-8

Dickson RB, Lippman ME (eds): Genes, Oncogenes and Hormones: Advances in Cellular and Molecular Biology of Breast Cancer. 1992. ISBN 0-7923-1748-3

Humphrey G, Bennett, Schraffordt Koops H, Molenaar WM, Postma A (eds): Osteosarcoma in Adolescents and Young Adults: New Developments and Controversies. 1993. ISBN 0-7923-1905-2

Benz CC, Liu ET (eds): Oncogenes and Tumor Suppressor Genes in Human Malignancies. 1993. ISBN 0-7923-1960-5.

Freireich EJ, Kantarjian H (eds): Leukemia: Advances in Research and Treatment. 1993. ISBN 0-7923-1967-2

Dana BW (ed): Malignant Lymphomas, Including Hodgkin's Disease: Diagnosis, Management, and Special Problems. 1993. ISBN 0-7923-2171-5

Nathanson L (ed): Current Research and Clinical Management of Melanoma. 1993. ISBN 0-7923-2152-9

Verweij J, Pinedo HM, Suit HD (eds): Multidisciplinary Treatment of Soft Tissue Sarcomas. 1993. ISBN 0-7923-2183-9

Rosen ST, Kuzel TM (eds): Immunoconjugate Therapy of Hematologic Malignancies. 1993. ISBN 0-7923-2270-3

Sugarbaker PH (ed): Hepatobiliary Cancer. 1994. ISBN 0-7923-2501-X

Rothenberg ML (ed): Gynecologic Oncology: Controversies and New Developments. 1994. ISBN 0-7923-2634-2

Dickson RB, Lippman ME (eds): Mammary Tumorigenesis and Malignant Progression. 1994. ISBN 0-7923-2647-4

Hansen HH (ed): Lung Cancer. Advances in Basic and Clinical Research. 1994. ISBN 0-7923-2835-3

Goldstein LJ, Ozols RF (eds): Anticancer Drug Resistance. Advances in Molecular and Clinical Research. 1994. ISBN 0-7923-2836-1

Hong WK, Weber RS (eds): Head and Neck Cancer. Basic and Clinical Aspects. 1994. ISBN 0-7923-3015-3

Recent Advances in Clinical Trial Design and Analysis

edited by

Peter F. Thall, Ph.D.
Department of Biomathematics
The University of Texas M.D. Anderson Cancer Center
Houston, Texas

SPRINGER SCIENCE+BUSINESS MEDIA, LLC

Library of Congress Cataloging-in-Publication Data

Recent advances in clinical trial design and analysis / edited by
Peter F. Thall.
p. cm. — (Cancer treatment and research; CTAR 75)
Includes bibliographical references and index.
ISBN 978-1-4613-5830-5 ISBN 978-1-4615-2009-2 (eBook)
DOI 10.1007/978-1-4615-2009-2
1. Clinical trials. 2. Medicine — Research — Statistical methods.
I. Thall, Peter F. II. Series: Cancer treatment and research; v. 75.
[DNLM: 1. Clinical Trials — methods. 2. Research Design.
3. Statistics. W1 CA693 v. 75 1995 / W 20.5 R295 1995]
R853.C55R426 1995
619 — dc20
DNLM/DLC
for Library of Congress 94-41111
CIP

Printed on acid-free paper.

Contents

List of Contributors

Jeffrey M. Albert, National Institutes of Health, Solar Building, Room 2B26, Bethesda, MD 20892, (for overnight: 6003 Executive Boulevard, Room 2B21, Rockville, MD 20852)

Oliver M. Bautista, The Biostatistics Center, Department of Statistics/ Statistical Computing, The George Washington University, 6110 Executive Boulevard, Rockville, MD 20852

Donald A. Berry, Institute of Statistics and Decision Sciences, and Cancer Center Biostatistics, P.O. Box 90251, Duke University, Durham, NC 27706

John Crowley, Fred Hutchinson Cancer Research Center Seattle, WA 98104

David L. DeMets, Department of Statistics University of Wisconsin–Madison, 1210 W. Dayton Street, Madison, WI 53706

Dennis O. Dixon, National Institutes of Health, Solar Building, Room 2B21, Bethesda, MD 20892, (for overnight: 6003 Executive Boulevard, Room 2B21, Rockville, MD 20852)

Boris Freidlin, Emmes Corporation, 11325 Seven Locks Road, Potomac, MD 20854

Richard D. Gelber, Division of Biostatistics, Dana Faber Cancer Institute, 44 Binney Street, Boston, MA 02155

Shari Gelber, Division of Biostatistics, Dana Faber Cancer Institute, 44 Binney Street, Boston, MA 02155

Patricia M. Grambsch, Division of Biostatistics, School of Public Health, University of Minnesota, A-460 Mayo Bldg., Box 197, 420 Delaware Street SE, Minneapolis, MN 55455

John M. Lachin, The Biostatistics Center, Department of Statistics/Statistical Computing, The George Washington University, 6110 Executive Boulevard, Suite 750, Rockville, MD 20852

Gordon Lan, The Biostatistics Center, Department of Statistics/Statistical Computing, The George Washington University, 6110 Executive Boulevard, Suite 750, Rockville, MD 20852

Michael LeBlanc, Department of Preventive Medicine and Biostatistics, University of Toronto Toronto, M5S 1A8 Ontario, CANADA

D.Y. Lin, Department of Biostatistics, SC-32, University of Washington, Seattle, WA 98195

Cyrus R. Mehta, Department of Biostatistics, Harvard School of Public Health and, Cytel Software Corporation, 675 Massachusetts Avenue, Cambridge, MA 02139

Richard M. Simon, Chief, Biometrics Research Branch, Cancer Therapy Evaluation Program (CTEP), National Cancer Institute (NCI), EPN — Room 739, Bethesda, MD 20892

Peter F. Thall, Department of Biomathematics, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Box 237, Houston, TX 77030

Preface

In early 1993, E.J. Freireich, M.D., asked me to edit a book on statistical methods in clinical trials. Initially, I felt my job would be to assemble a group of biostatisticians, each willing to contribute a chapter to the book. The goal was to provide descriptions of some of the more modern developments in statistical methods for clinical trials, for both medical and biostatistical readers. Dr. Freireich's request presented a challenge to me, because the proposed book was to appear in a series that previously had consisted entirely of medical papers directed at a medical audience. I chose biostatisticians currently doing cutting-edge research and asked each to write a survey, possibly including new research, with a combined medical–biostatistical audience in mind. The collection of chapters in this book is the result.

Scientists tend to communicate within their own areas of expertise, typically using a specialized language that often seems incomprehensible to those outside their field. The problem of interdiscipline communication — communicating ideas between distinct groups of people who think differently and use different languages — is one of the most difficult and important problems in science. When one of the fields is statistics, the problem can be even more difficult and challenging than usual. It is more difficult because statistical concepts have their foundation in mathematics, in particular probability, and translating mathematical concepts into language understandable by nonmathematicians is, in my experience, one of the most intellectually demanding activities that can be undertaken.

Communicating statistical concepts to biomedical scientists who have had little or no statistical training is critically important, however. Clinical trials have two purposes — to treat the patients in the trial, and to obtain information that increases our understanding of the disease and especially of how patients respond to treatment. Statistical design provides a means to achieve both these aims, while statistical data analysis provides methods for extracting useful information from the trial data.

Recent advances in statistical computing, both in the computers themselves and in conceptual devices such as computational algorithms, have enabled statisticians to implement very rapidly a broad array of methods that previously were either impractical or impossible. Biostatisticians thus have become able to provide much greater support to medical researchers working in both clinical and laboratory settings. As our collective toolkit of techniques for analyzing data and designing clinical trials and laboratory studies has grown, however, it has become increasingly difficult for each of us biostatisticians to keep up with all the developments in our own field. The task of communicating these advances to our medical colleagues thus has become doubly difficult just as we are entering a new age in which so many truly powerful statistical methods are becoming practical realities.

This book is one attempt, among many currently ongoing, to explain some of the more recent developments in biostatistics to clinicians and scientists who work in clinical trials. Each of the chapters describes a very recent development in statistical methodology that is a powerful tool for planning or analysis. The chapters are written at a variety of technical levels, reflecting the individual styles of the contributors. Thus, I think unavoidably, some chapters will be more accessible than others. If these chapters make you aware of even one new method that you find useful, however, then I have achieved my goal.

Peter F. Thall

1. The alpha spending function approach to interim data analyses

David L. DeMets and Gordon Lan

Introduction

Over the past three decades, clinical trials have become one of the major standards for evaluating new therapies and interventions in medicine [1–3]. Numerous clinical trials have been conducted during this period across a wide variety of diseases, evaluating drugs, procedures, devices, and biologic materials. The fundamentals of the design, conduct, and analyses of clinical trials have been developed and refined during this period as well. One such fundamental is that clinical data should be carefully monitored during the course of the trial so that unexpected or unacceptable toxicity can be detected as soon as possible in order to minimize patient exposure; in addition, trials should not be continued longer than necessary to prove the benefits of the therapy or intervention under study, or to understand the trade-offs between the benefits and risks of the therapy. In order to accomplish this goal, the National Institutes of Health sponsored a committee in the 1960s to develop guidelines for the conduct of clinical trials. The chair of this committee was Dr. Bernard Greenberg from the University of North Carolina, and the report, which was issued in 1967, has become known as the Greenberg Report [4], although it was only recently published in the literature. This report endorses the concept of interim review of data by an independent Data and Safety Monitoring Board (DSMB), a committee that has no conflict of interest for the study. This typically means that committee members should not be investigators entering patients into the trial. The Coronary Drug Project (CDP) [5] was one of the first trials to implement the Greenberg model.

The decision to terminate a trial early due to unacceptable toxicity or substantial and convincing evidence of benefit is complex and must account for many factors [5–17]. These include possible imbalances in risk factors between treatment groups, whether the patients have the risk profile assumed in the design, patient compliance to therapy, quality and timeliness of data, possible sources of bias, consistency of primary and secondary outcome variables, the benefit-to-risk ratio, consistency of results with external data, and the impact of early termination on the medical community as well as the

Peter F. Thall (ed.), RECENT ADVANCES IN CLINICAL TRIAL DESIGN AND ANALYSIS.
Copyright © 1995 Kluwer Academic Publishers, Boston. All rights reserved. ISBN 978-1-4613-5830-5.

public. Evaluation of these issues goes beyond routine statistical tests and requires the collective judgment of experts, such as those represented by a DSMB. That is, the DSMB usually has members with clinical, laboratory, statistical, and epidemiological expertise and often someone with a background in ethics related to patient research.

One of the issues identified in the CDP experience was that repeated analysis of accumulating data raises the chances of false-positive claims if standard statistical methods are used at each analysis with no adjustments for the repetition. The problem of repeated or sequential testing of data was already well known by that time due to previous or ongoing work [18–25]. Canner [25] describes some of the methodology used in the CDP interim analyses. While many statistical methods were available, the CDP experience clearly indicated that the decision-making process to terminate a trial early due to evidence of toxicity or benefit is complex, and statistical methods alone are not sufficient [5]. Several trials conducted since then have confirmed this principle [6–12].

Nevertheless, while statistical methods cannot be used as termination rules, they can be very helpful as termination or stopping guidelines [8–17].

A great deal of statistical research has occurred during the past 15 years to develop, adapt, or modify existing statistical methods in order to provide better tools for this complex decision process, including a recent text by Whitehead [26]. This research has spanned frequentist, Bayesian, and decision-theoretic points of view. We shall focus our attention on the frequentist viewpoint. In particular, the frequentist approach attempts to minimize false-positive claims by controlling the type I error probability or 'alpha level.' Haybittle [27] and Canner [25] introduced the idea of using a very conservative criterion at each interim analysis. Work by Pocock [28] and O'Brien and Fleming [29] introduced an approach referred to as 'group sequential' analyses of interim data, which can be viewed as an extension of work pioneered by Armitage and others [22] on repeated significance testing. The Pocock modification focused on the idea that when the DSMB meets periodically, an additional group of subjects or events has been observed. The number of interim analyses must be specified in advance, and the number of patients or events must be divided equally between analyses. However, how many times or exactly when a DSMB might meet to conduct the safety and benefit assessment is not always easy to predict or prescribe exactly. For example, the number of events observed between successive meetings of the DSMB typically vary, i.e., are not equal. Moreover, the DSMB might spot a worrisome trend and request additional meetings.

Lan and DeMets [30] extended the group sequential concept to a very flexible method that controls the overall alpha level while allowing for the number and exact timing of the interim analyses to remain unspecified a priori. This general approach which has been used in a number of clinical trials, will be described here. This chapter is an expanded version of summary papers published previously [31–35], and it also summarizes numerous other papers on this topic.

The alpha spending function concept

In fixed-sample, classical nonsequential designs, the allowed alpha level corresponds to a single, final analysis. However, in repeated interim analyses, the cumulative Type I error rate increases with each interim evaluation. Armitage, McPherson, and Rowe [22] provided quantitative results showing the actual cumulative type I error for various numbers of interim analyses while using the conventional fixed-sample critical values each time. For example, if the conventional critical value of 1.96 is used, corresponding to a fixed-sample two-sided 0.05 significance level, the actual type I error rate is nearly 0.15 for five interim analyses and almost 0.20 for 10 analyses. Five to ten interim analyses are not uncommon for larger, longer-term follow-up trials, but clearly type I error rates of 15% to 20% are unacceptably high for critical or pivotal clinical trials.

The goal of the general group sequential approaches [28–30] is to control the type I error rate. The alpha spending function, which will be formally defined in the next section, allocates some of the prespecified type I error to each interim analyses. The specific models proposed by Pocock [28] and O'Brien and Fleming [29] are special cases of this approach. The alpha spending function allocates the total allowable type I error rate through a function based on the information accrued during the trial, such as the total number of observed patients or events. That is, the spending function depends on the fraction of patients or events observed at a particular interim analysis out of the total number of patients or events expected or designed for. This fraction, t^*, referred to as the information fraction, indicates how much of the trial has been completed in terms of the accumulated information, and thus indicates how much of the allowable type I error rate should be allocated. The value of the information fraction must be between 0 and 1. The alpha spending function must be equal to 0 at $t^* = 0$ and equal to alpha at $t^* = 1.0$, and it is nondecreasing in between. An example of a spending function is given in figure 1 for a spending function that corresponds approximately to an O'Brien–Fleming group sequential model. For each interim analyses, the allocated type I error is determined by the alpha spending function, which in turn corresponds to an adjusted critical value for the test statistic computed at that analysis.

One limitation of previous group sequential methods is the requirement that both the number and the exact time be specified in advance. For example, a trial design might specify that five interim analyses are planned at information fractions 0.20, 0.40, 0.60, 0.80, and 1.0. However, as the trial

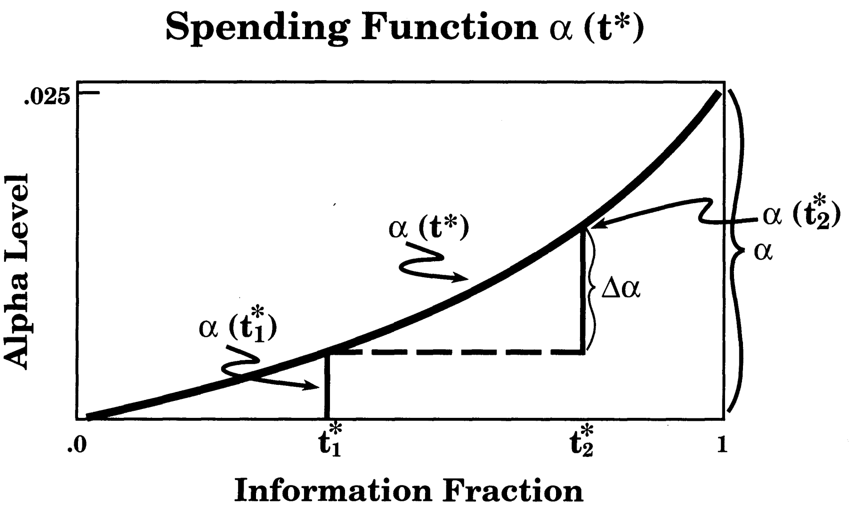


Figure 1. Alpha spending function indicating additional type I error rate, Δα, allocated between interim analyses t_1^* and t_2^*.

progresses, the DSMB may not be able to meet when the information fraction is exactly those prespecified fractions, or may need to meet more frequently due to emerging toxicity or beneficial trends. This issue was raised by one of the early cardiovascular trials that used the O'Brien–Fleming group sequential model [6,10]. However, the alpha spending function does not require either a specific number of interim analyses or specific times when they must occur. It does however require that the particular spending function be specified in advance and that we know how many total patients or events to expect in the trial. That is, the trial sample size must be specified in advance, which most well-designed trials will in fact require. Details regarding this flexible alpha spending function will be described in the remainder of this chapter. When the number of patients or the number of events for the whole study is uncertain, we need some modifications to apply the spending function approach. This is illustrated below in the section on survival analysis.

Formal alpha spending function $\alpha(t)$

Since the Lan–DeMets alpha spending function approach was introduced in 1983 [30], a decade of research on this flexible method has emerged, indicating how it can be used in a variety of settings. These include comparison of proportions, means, survival curves, and repeated measures, as well as methods for computing confidence intervals and *p*-values. In the following sections, we shall first formally define the alpha spending function

[30,33–37] and discuss issues related to it and then illustrate the design and analysis methods listed above.

Definition

In the fixed sample setting, we often wish to evaluate the null hypothesis of no treatment effect using a test statistic Z compared to a critical value Z_C that corresponds to a prespecified type I error or alpha level. We shall consider only two-sided symmetric sequential tests, but extensions to one-sided or asymmetric tests are self-evident and straightforward [38,39]. A more theoretical development of this approach may be found in Lan and Zucker [35]. The group sequential method for interim analyses defines a critical value for each analysis $Z_C(k)$, $k = 1, 2, \ldots, K$, such that the overall type I error rate is maintained. In the Lan and DeMets [30] approach, the total type I error is allocated to each analysis through the spending function, which in turn determines the value of $Z_C(k)$. The trial continues to accrue patients or events if, at the kth interim analysis,

$$|Z(k)| < Z_C(k) \quad \text{for} \quad k = 1, 2, \ldots, K-1,$$

where $Z(k)$ is the test statistic for the kth analysis. If the test statistic exceeds the boundary or critical value, then early termination of the trial should be considered, after careful consideration of all the evidence as discussed above. At the final planned analysis, the null hypothesis of no treatment difference would be accepted if $|Z(K)| < Z_C(K)$. The null hypothesis would be rejected if the test statistic $|Z(K)| \geqslant Z_C(K)$.

The test statistic $Z(k)$ for all the groups at the kth interim analysis is obtained from a summary of the results from each of the previous k groups; that is,

$$Z(k) = \{\sqrt{I_1}\,Z^*(1) + \sqrt{I_2}\,Z^*(2) + \ldots + \sqrt{I_k}\,Z^*(k)\}/\sqrt{(I_1 + I_2 + \ldots + I_k)}\,,$$

where I_i and $Z^*(i)$ respectively represent the amount of information and the summary statistic for the ith group, which is comprised of the data points accumulated between the $(i) - 1$th and ith DSMB meetings.

Consider the clinical trial to be completed in calendar time t between $[0,T]$, where T is the scheduled recruitment time for an immediate-response study or follow-up time for an event-based study [34,40,41]. During the trial at calendar time t, let t^* denote the information fraction, which is the observed information divided by the total information expected or designed for. If at the kth interim analysis, we observe information $I_1 + I_2 + \ldots + I_k$ and expect to have total information of I at the scheduled end of the study, then the information fraction t_k^* at calendar time t_k^* is $(I_1 + I_2 + \ldots + I_k)/I$. For comparing means or proportions, t_k^* is approximated by n/N, the observed sample size divided by the expected maximum sample size N. For survival studies, t_k^* is approximated by d/D, the number of observed deaths divided by the total expected number of deaths D. The information fraction will be more formally defined in the next section. The Lan and DeMets alpha spending function, $\alpha(t^*)$ is defined such that $\alpha(0) = 0$ and $\alpha(1) = \alpha$. Group sequential boundaries or critical values for the test statistic computed at the kth interim analysis can be determined according to the spending function $\alpha(t^*)$. Let analysis times and information times be defined such that $0 < t_1 < t_2 < \ldots < t_K \leqslant T$ and $0 < t_1^* < t_2^* \ldots < t_k^* = 1$, where K denotes the last and final analysis. Then we can determine the boundary values $Z_C(k)$ at t_k for $\alpha(t_k^*)$ by solving successively, under the null hypothesis of no treatment effect,

$$\begin{aligned} &P_0\{|Z(1)| \geqslant Z_C(1) \\ \text{or}\quad &|Z(2)| \geqslant |Z_C(2)| \text{ or} \ldots \text{or } |Z(k)| \geqslant Z_C(k)\} = \alpha(t_k^*) \end{aligned}$$

for a two-sided test of the hypothesis. Note that $Z_C(k)$ is determined by the spending function and the information fractions $t_1^*, t_2^*, \ldots t_k^*$, but does not depend on future information fractions or on the value of K.

The increment $\alpha(t_k^*) - \alpha(t_{k-1}^*)$ represents the additional type I error rate or alpha level that is allocated to the kth interim analysis. For a single fixed-sample design,

$$P_0\{|Z(K = 1)| > Z_C(K = 1)\} = \alpha(1) = \alpha.$$

That is, the total alpha is spent all at once at the end of the trial. By examining the data at various intervals of information, we allocate the total alpha to each analyses such that

$$\Sigma_k\{\alpha(t_k^*) - \alpha(t_{k-1}^*)\} = \alpha, \quad k = 1, 2, \ldots, K.$$

Evaluation of the probability P_0 requires knowing the distribution of the sequence of test statistics $\{Z(1), Z(2), \ldots, Z(k)\}$ under the null hypothesis. If each group statistic $Z^*(i)$, $i = 1, 2, \ldots, k$ is normal with mean zero and unit variance and if they are independent, then the summary statistic $Z(k)$ also has a normal distribution with mean zero and unit variance. For this case, the distribution function has a special form as a recursive density function that can be numerically integrated to obtain the value of the type I error rate spent up to that point for a given set of critical values [22,28,30]. If the individual group statistics do not have this independent increment structure but still have some known or approximated multivariate distribution, the spending-function approach can still be implemented, but it is somewhat more complicated. Fortunately, most of the common applications have this independent increment structure, as will be described below in the section on applications.

The Pocock [28] boundary corresponds to a constant critical value for each interim analysis, $Z_C(k) = Z_P$. The O'Brien – Fleming [29] boundary decreases in absolute value as the information fraction increases such that $Z_C(k) = Z_{OBF}/\sqrt{(n/N)} = Z_{OBF}/\sqrt{(t^*)}$, where Z_{OBF} is a constant value.

Spending functions can be defined that approximate O'Brien–Fleming [29] or Pocock [28] boundaries, or something in between [30], as follows:

$$\begin{array}{ll} \alpha_1(t^*) = 2 - 2\Phi(Z_{\alpha/2}/\sqrt{(t^*)}) & \text{O'Brien–Fleming type} \\ \alpha_2(t^*) = \alpha \cdot \ln(1 + (e - 1)^{t^*}) & \text{Pocock type} \\ \alpha_3(t^*) = \alpha \cdot t^* & \text{Uniform} \end{array}$$

where Φ denotes the standard normal cumulative distribution function. The shape of the spending functions for these three functions are shown in figure 2 with an overall 0.05 type I error rate — for example, 0.025 allocated to a positive trend and 0.025 to a negative trend.

In table 1, we have indicated the comparison of the critical values or monitoring boundaries for the test statistic computed in this manner to those provided in the Pocock [28] and O'Brien–Fleming [29] papers for a total of $K = 5$ analyses at equally spaced information fractions $t^* = 0.2, 0.4, 0.6, 0.8$, and 1.0. Note that the boundaries are not exactly equivalent, since they are defined differently, but they are very close. Pocock's method yields a constant critical value of 2.41 in comparison to a naive boundary value of 1.96. The O'Brien–Fleming coefficient is 2.04, which provides the critical values when adjusted by the information fraction. It should be emphasized that these two methods initially required equally spaced increments of information, with the number of interim analyses to be specified in advance. The Lan–DeMets version does not have these constraints. The boundaries for α_1

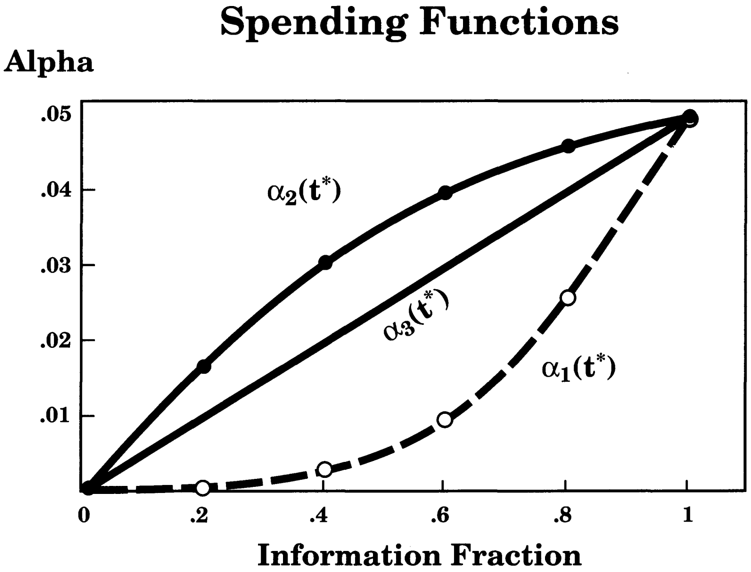


Figure 2. Comparison of spending functions $\alpha_1(t^*)$, $\alpha_2(t^*)$, and $\alpha_3(t^*)$ at information fractions $t^* = 0.2, 0.4, 0.6, 0.8$, and 1.0.

Table 1. Comparison of boundaries using spending functions with Pocock (P) and O'Brien–Fleming (OBF) methods (α = 0.05, t^* = 0.2, 0.4, 0.6, 0.8, 1.0)

t^*	$\alpha_1(t^*)$	OBF	$\alpha_2(t^*)$	P
0.2	4.90	4.56	2.44	2.41
0.4	3.35	3.23	2.43	2.41
0.6	2.68	2.63	2.41	2.41
0.8	2.29	2.28	2.40	2.41
1.0	2.03	2.04	2.39	2.41

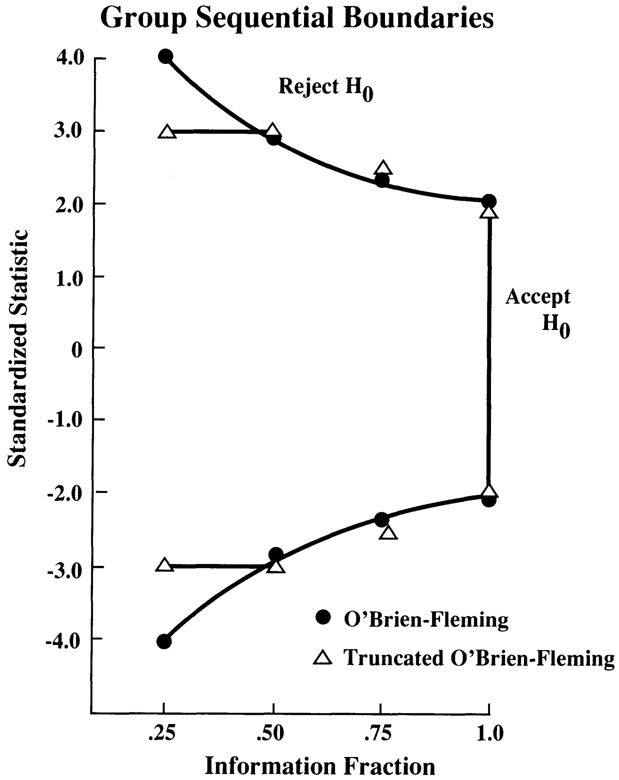


Figure 3. Upper boundary values corresponding to the $\alpha_1(t^*)$ spending function for $\alpha - 0.05$ at information fraction t^* = 0.25, 0.50, 0.75, and 1.0 and for a truncated version at a critical value of 3.0.

are shown in figure 3 for interim analyses at t^* = 0.25, 0.50, 0.75, and 1.0. Since the early boundary values may be very extreme, we can also truncate these extreme boundaries at some large value such as ± 3.0 without affecting the rest of the boundary. More general classes of spending functions have also been developed [36,37], but the three spending functions described here represent the range of alternatives. Nonsymmetric boundaries are also possible [37–39] by setting different alpha levels to be spent for positive or negative treatment effects. The generalizations from the methods described here are straightforward.

Information fraction

A simple way to describe 'statistical information' is that each patient randomized in a clinical trial contributes up to one unit of 'statistical information' to a specific endpoint [40–42]. When the data are analyzed, a patient's contribution is 'one' if his or her endpoint has been completely measured and 'less than one' if it was only partially measured. The exact amount of information contributed by a patient depends on the nature of the endpoint, the patient's follow-up time, the statistical test used for treatment group comparisons, and possibly some other factors. Let us use some examples to elaborate on this point. Suppose 'one-week mortality' is the endpoint being considered. A patient, one week after randomization, is either dead or alive and contributes one unit of information to the study. Another patient, three days after randomization, can also contribute one unit of information if he or she is dead by then. If still alive, he or she contributes no information to one-week mortality. If an endpoint can be measured completely soon after a patient enters a study, then 'the amount of information observed' practically has the same meaning as 'the number of patient randomized into the study.' Other examples of immediate outcomes are 24-hour blood pressure change or 90-minute reperfusion rate.

The situation is more complicated when we are interested in the exact survival time of the patients [40,41]. In most clinical trials, we do not follow all the patients until their deaths before we analyze data. Therefore, the amount of information available when the data are analyzed is usually less than the number of patients in the study. Obviously, the longer the follow-up, the more information a patient contributes to the study. Similarly, if we are interested in, say, the change of FEV_1 (forced expiratory volume in one second) of lung disease patients, we do not measure the patients' lung functions continuously. Instead, we ask the patients to come back for periodic checkups and take their FEV_1 measures then. In this case, the amount of information a patient contributes depends on the frequency and spacings between visits of this specific patient.

When we design a two-group clinical trial, we assume a specific treatment difference and then compute the amount of information required to reach a certain power. If the data are analyzed only once after all the information has been accumulated, we have only one chance to make a type I error or to make a false-positive decision. We consider this design to be 'spending' all the alpha at the end of the study. In many large-scale clinical trials, data are monitored periodically, and decisions on treatment comparisons are made sequentially [42]. In order to maintain the overall alpha at a desired level, we 'spend' this fixed amount of alpha as information is being accumulated. In other words, the 'spending function' specifies the proportion of alpha to be spent as a function of 'information accrued.' However, the total information varies from study to study, and it is more convenient to express the spending of alpha as a function of the information fraction

$$t^* = \frac{\begin{array}{c}\text{(amount of information contributed by all the}\\ \text{patients when data are analyzed)}\end{array}}{\text{(amount of total information for the study)}},$$

which varies from 0 to 1 as the study proceeds.

Since the total amount of alpha is fixed for a study, a conservative monitoring plan would spend a small amount of alpha at the beginning of the study so that more can be reserved for the later part of the study. Conversely, an aggressive monitoring scheme spends a large amount of alpha at the beginning and leaves a small amount of alpha for later, such as $\alpha_2(t^*)$. Note that this general concept can be visualized as the 'shape' of the spending function. Roughly speaking, an aggressive monitoring spending function results in earlier stopping when a treatment difference is large, but has less power to detect a treatment difference when compared to a more conservative spending function. Conservative spending functions do not easily allow for early termination, and their final critical value is close to the fixed-sample critical value as in $\alpha_1(t^*)$. If the amount of total information is uncertain, but the duration of the trial is fixed, then the information fraction at the time of data monitoring has to be estimated. An illustration is given below in the section on survival analysis.

Change of frequency and overruling

The methods initially proposed by Pocock [28] and O'Brien and Fleming [29] assumed that the number and timing of the interim analyses are fixed in advance. While most of the information can be captured in a few interim analyses [12,13,37,43], the DSMB may request additional interim analyses due to emerging trends. When the alpha spending function approach was introduced by Lan and DeMets [30], concern was expressed that this very flexible approach could be abused if the frequency of interim analyses were changed due to emerging trends. This concern was addressed by several researchers, including Lan and DeMets [44] and Proschan et al. [45]. Lan and DeMets simulated several scenarios in which the frequency of interim analyses would be doubled if the emerging trends got to within 80% of the current critical value or boundary. The results of one simulation study are given in table 2 and, as shown, there is a negligible increase in the type I error. Proschan et al. [45] considered more intense strategies to abuse the spending function. In their worst case, the alpha level or type I error rate was doubled, but for the more common spending functions, such as the

Table 2. Simulation results for impact of changing frequency on the alpha level and power

	Spending function			
	$\alpha_1(t^*)$ — O'Brien–Fleming		$\alpha_2(t^*)$ — Pocock	
θ	Rule 1	Rule 2	Rule 1	Rule 2
0	0.024	0.025	0.025	0.026
2	0.508	0.511	0.431	0.432
4	0.845	0.846	0.782	0.782
5	0.976	0.976	0.960	0.959

Rule 1: Interim analysis at $t^* = 0.25, 0.50, 0.75$.
Rule 2: If test statistic at interim analysis is within 80% of a boundary per rule 1, double the frequency of interim analyses.
$\theta = \Delta\sqrt{K}$, where Δ is the noncentrality parameter of the test statistic.

O'Brien–Fleming type spending function, the alpha level did not inflate noticeably. It is, of course, not permissible to change the spending function during the course of the trial. This point needs to be emphasized because, if spending functions are changed, there is no longer any control over type I error, and serious abuse is possible — that is, the interim monitoring process would have little credibility.

Application: design and analysis

So far, we have described the alpha spending function in terms of a general test statistic Z evaluating a treatment effect. In this section, we shall describe how this general approach can be applied to a few specific test statistics. For each case, we shall describe the test statistic, the design approach, and the implementation for the interim analysis.

Although the alpha spending function provides the desired flexibility in the analysis phase, for design purposes, it does require some prior specification of the number of interim analyses and the times. Once the design, including the target sample size, has been established, the frequency and timing of the interim analyses may vary using the alpha spending approach without any significant impact on the overall type I error rate [44]. Thus, the design strategy or alpha spending function [46] is essentially the same as that described by Pocock [28].

Comparison of means

Some clinical trials compare mean levels of response. The basic hypothesis being tested is that there is no difference in mean values, μ (i.e., no treatment effect). We have some treatment effect in mind that we would like to detect (the alternative hypothesis). More formally, let the null hypothesis H_0 be defined

$$\begin{aligned} H_0&: \quad \mu_C - \mu_T = 0 \\ H_A&: \quad \mu_C - \mu_T = \delta \neq 0 \end{aligned}$$

where μ_C and μ_T represent the true control and treatment group means, and δ represents the value of hypothesized treatment effect compared to a control. We would obtain a sample mean from each group, control and treatment, and then compare means as follows:

$$Z = \frac{\bar{X}_C - \bar{X}_T}{\sigma\sqrt{1/m + 1/m}}$$

assuming equal sample size m for each group for simplicity, where σ denotes the population standard deviation. Since σ is unknown, we may estimate σ by $\hat{\sigma}$, the sample standard deviation. For a large enough sample size, this statistic has approximately a normal distribution with mean 0 and unit variance under the null hypothesis. Under the alternative hypothesis ($\delta \neq 0$), this statistic has a normal distribution with mean Δ and unit variance, where

$$\Delta = (\mu_C - \mu_T)/(\sigma\sqrt{1/m + 1\ 1/m}).$$

In the design phase, we might specify a total of K planned interim analyses after every increment of n patients per group. The test statistic after the kth such group is

$$Z_k = \frac{\bar{X}_C - \bar{X}_T}{\sqrt{2\sigma^2/nk}} \quad k = 1, 2, \ldots, K,$$

where $\bar{X}_C$ and $\bar{X}_T$ are the means across all k groups.

For this case, we can write the value of the parameter Δ, the expected value of the statistic under the alternative hypothesis, as

$$\Delta = \sqrt{n}(\mu_C - \mu_T)/\sqrt{2\sigma^2} = \sqrt{n}\delta/\sqrt{2\sigma^2}$$

so that

$$n = \frac{2\Delta^2\sigma^2}{(\mu_C - \mu_T)^2} = 2\Delta^2\sigma^2/\delta^2.$$

In order to design our studies, we evaluate the previous equation for n, the sample size per treatment per sequential group. Since the plan is to have K groups each of size $2n$, the total sample size $2N$ equals $2nK$. Now, in order to obtain the sample size in the context of the alpha spending function, we proceed as follows:

1. Fix the number of planned interim analyses K at equally spaced incre-

ments of information (i.e., $2n$ subjects). It is also possible to specify unequal increments, but equally spaced is sufficient for design purposes.

2. Obtain the boundary values of the K interim analyses under the null hypothesis H_0 to achieve a prespecified overall alpha level, α, for a specific spending function $\alpha(t^*)$.
3. For the boundary obtained, obtain the value of Δ to achieve a desired power $(1 - \beta)$.
4. Determine the value of n that determines the total sample size $2N = 2nK$.
5. Having computed these design parameters, one may conduct the trial with interim analysis to be done based on the information fraction t_k^* approximated by

   $$t_k^* = \text{Number of subjects observed}/2N$$

   at the kth analysis. The number of actual interim analyses may not be equal to K, but the alpha level and the power will be affected only slightly [46].

As a specific example, consider using an O'Brien–Fleming-type alpha spending function $\alpha_1(t^*)$ with a two-sided 0.05 alpha level and 0.90 power. We wish to test an alternative hypothesis H_A: $\mu_C - \mu_T = \delta = 0.5\sigma$, a difference of half a standard deviation. We also plan to perform five ($K = 5$) interim analyses at $t^* = 0.2, 0.4, 0.6, 0.8$, and 1.0. Using previous publications [29] or available computer software, we obtain boundary values 4.56, 3.23, 2.63, 2.28, and 2.04. Using these boundary values and available software, we again find that $\Delta = 1.28$ provides the desired power of 0.90 to detect $\delta = 0.5\sigma$. Thus, substituting for Δ, we find

$$n = \frac{2\Delta^2\sigma^2}{(\frac{1}{2}\sigma)^2} = 8(1.28)^2 \simeq 13.1;$$

i.e., we would require a total sample size of $2N = 2nk = 2(13)5 = 130$ patients.

As we conduct the kth interim analysis, we will compute the exact group sequential boundary $Z_C(k)$ through the use function $\alpha_1(t_k^*)$, where the information fraction t_k^* will be approximated by the observed sample size divided by 130.

Comparison of proportions

Many trials also compare the frequency of events between two treatment groups. The process for design and interim analyses proceeds in a similar fashion to that described for the comparison of means. Here,

$$\begin{aligned} &\mathrm{H}_0: && p_C - p_T = 0 \\ &H_A: && p_C - p_T = \delta \neq 0 \end{aligned}$$

where p_C and p_T denote the unknown response rates in the control and new-treatment groups, respectively. We would estimate the unknown parameter by $\hat{p}_C$ and $\hat{p}_T$, the observed event rates in our trial. For a reasonably large sample size, we often use the following test statistics:

$$Z = \frac{\hat{p}_C - \hat{p}_T}{\sqrt{\hat{p}(1 - \hat{p})(1/m_C + 1/m_T)}}$$

to compare event rates where $\hat{p}$ is the combined event rate across treatment groups. For sufficiently large m_C and m_T, this statistic has an approximate standard normal distribution with mean Δ and unit variance under the null hypothesis H_0: $\Delta = 0$. In this case,

$$\Delta = \sqrt{n}(p_C - p_T)/\sqrt{2p(1 - p)} = \sqrt{n}\delta/\sqrt{2p(1 - p)}$$

and

$$n = \frac{2\Delta^2 p(1 - p)}{\delta^2}.$$

Similarly to the example given above in the section on comparison of means, we might design a trial for $K = 5$ interim analyses using an O'Brien–Fleming-type spending function $\alpha_1(t^*)$ at equally spaced increments for a two-sided alpha level of 0.05. If we specify $p_C = 0.6$, $p_T = 0.4$ ($p = 0.5$) under the alternative hypothesis, then we can obtain a sample size as follows. For $\Delta = 1.28$,

$$n = \frac{2(1.28)^2(0.5)(0.5)}{(0.2)^2} = 20.5,$$

and we have a total sample size of 2(21)5 = 210 subjects. We can then proceed to conduct interim analysis times at information fraction t_k^* equal to the observed number of subjects divided by 210.

Survival analysis

In survival analysis, linear rank statistics, which include the logrank statistic and the Wilcoxon statistic, are commonly used for two-group comparisons of survival curves [47]. The logrank statistic takes the form $\Sigma_i(O_i - E_i)$, where the sum is over all the events. The observed value O_i indicates whether the ith event comes from group 1. To be more specific, $O_i = 1$ (or 0) if the ith event is for a group 1 (group 2) patient, respectively. The expected value E_i corresponds to the proportion of group 1 patients at risk when the ith event occurs. A linear rank statistic takes the form $\Sigma_i W_i(O_i - E_i)$, where W_1 corresponds to the 'weight' of the ith event. For the logrank test, the weights are equal to 1. The Wilcoxon statistic, for example, puts more weight on earlier events than later events [48]. In the 1980s there were some important developments in group sequential monitoring of survival data [49–56]. It has been shown that the sequential methods developed for the comparison of two means also apply to the comparison of two survival distributions. The concept of information, however, needs some modification. The term *information* corresponds to the variance of the linear rank statistic; hence it has different interpretations for different tests. We use the logrank test, which is the logrank statistic normalized by its standard deviation, to illustrate the concept of information in the survival setting.

First of all, sample size alone is not enough to reflect the amount of information in a survival study. Suppose we plan to recruit 2000 patients, 1000 each in group 1 (standard treatment) and group 2 (new treatment). If each patient is followed for just one day, we may end up with few or even no events at all. Despite the large sample size, we will not have much information to distinguish the effects of the two treatments. In this setting, the information provided by each patient can be explained through the distribution function of the survival time. Consider the following hypothetical example:

Follow-up time	1 month	2 months	3 months	4 months
Probability of death	0.3	0.4	0.45	0.5

Suppose, at the first day of each month, we recruit 100 patients into the study. (A more realistic recruiting scenario will be discussed later.) After four months, we have recruited 400 patients. The first 100 patients in the study have been randomized for four months and we expect 100 × 0.5 = 50 events. The next 100 patients recruited have been in the study for three months, and we expect 100 × 0.45 = 45 events. Similarly, for the 100 patients recruited in the third and fourth months, we expect 100 × 0.4 = 40 and 100 × 0.3 = 30 events, respectively. The total expected number of events, 165 = 50 + 45 + 40 + 30, represents the amount of information available for the comparison of survival times for the two treatment groups. Note that the amount of information contributed by a patient depends on the 'time since randomization,' which is the duration between randomization and data analysis. A patient's contribution of information to the study is 0.3 after one month of randomization, 0.4 after two months, 0.45 after three months, and 0.5 after four months. The number 165 = (100 × 0.3) + (100 × 0.4) + (100 × 0.45) + (100 × 0.5) is the total of the contributions from the 400 patients in the study. In practice, we recruit patients every day; we need an extension of the above table to evaluate information accurately, but the fundamental principle is the same. When survival data are compared at calendar time t, the corresponding information fraction is

$$t^* = \frac{\text{(expected number of events by } t\text{)}}{\text{(expected number of events in the entire study)}}.$$

Since the expected number of events is not observable, we must use the observed number of events to replace them in practice. With this modification of information, the sequential boundaries presented earlier apply to the monitoring of the logrank test for comparisons of two survival curves. Note that the sequential boundaries are employed to control the type I error rate under the null hypothesis — namely, that there is no treatment difference. Under the alternative, we assume that the new treatment is 'better,' a term that in most practical situations is not very well defined [42]. However, under the proportional hazards model, the sequentially computed logrank statistics $\{Z_t\}$ behave like the $\{Z_t\}$ for the comparison of two means. That is, the methods involved in the design and data monitoring for the comparison of two means also apply to the comparison of two survival curves using the logrank test [49,50,53–57].

To design a study using the logrank test to compare the survival patterns of two treatment groups, the concept of information is expressed as the (expected) number of events, which corresponds to the number of patients in the comparison of two means. The treatment difference $(\mu_C - \mu_T)/\sigma$ in the comparison of means is replaced by log(hazard ratio) for the survival setting. If we can use a maximum information design, where the trial ends when a specified number of events is observed, then the evaluation of the information fraction at data monitoring is relatively straightforward. The information fraction may be estimated in one of three ways. We might estimate it by the fraction of observed control (placebo) group deaths of the total expected control group deaths. We might also compute the ratio of total observed deaths in both groups to that expected, where the expected number of deaths is estimated under the null hypothesis of no treatment difference. Alternatively, we might estimate the information fraction as the ratio of the total number of observed deaths to the total number of expected deaths, estimated under the alternative hypothesis. Any of these approaches is, valid, but the latter is preferred.

Due to budgeting or other logistical reasons, many studies are design to last for a specified period of calendar time. Such a design is called a maximum duration design, in contrast to a maximum information design. Here, we may not observe a prespecified number of events in the fixed time of follow-up. We could, of course, guess the total number of events to be observed, but we might over- or underestimate the number of expected events. For a maximum information design in a survival setting, several approaches to estimating the information fraction have been proposed [53]. One simple way is to estimate t^* by the fraction of study calendar time. Another more dynamic approach is to estimate the information fraction by the patient exposure time. For simplicity, we consider only the calendar time fraction estimate in this chapter.

We illustrate these methods for estimating t^* with data from the Beta-blocker Heart Attack Trial (BHAT) [6]. The BHAT [6] trial was a randomized double-blind multicenter trial evaluating the effectiveness of a beta-blocker drug, propranolol, in reducing mortality in patients who had recently suffered a myocardial infarction. With a two-sided significance level of 0.05 and a

90% power to detect a 20% reduction in mortality over a three-year follow-up, adjusting for noncompliance, a target sample size of 4000 patients was established. Recruitment was to be completed in two years and began in June of 1978. Follow-up was to end in June 1982. After a mean follow-up of almost two years, the trial was terminated nearly a year early due to a significant reduction in total mortality. Details of the decision process, given in [10], included the fact that the logrank statistic crossed the O'Brien–Fleming boundary. As already indicated, the numbers of deaths between analyses were not equal, and the frequency of analyses could have changed toward the end, although in fact it did not. The method we will present here was developed after the BHAT termination and does not reflect what actually happened. For our present purposes, we shall apply the O'Brien–Fleming-type spending function, $\alpha_1(t^*)$, with a two-sided 0.05 alpha level to monitor this trial in retrospect.

As indicated in table 3, BHAT was scheduled to be monitored seven times, each approximately six months apart. In practice, the BHAT trial was monitored at calendar times t_i = 11, 16, 21, 28, 34, and 40 months. BHAT was stopped early at t_6 = 40 (October 1981) with a logrank Z-value of $Z(6)$ = 2.82 favoring propranolol. The observed numbers of events at data monitoring were d_i = 56, 77, 126, 177, 247, and 318, respectively. The total number of events, D, expected at calendar time t_7 = 48 months (June 1982) was estimated to be 400, based on the lifetable available in October 1981 under the alternative assumption of a 20% reduction in mortality. The logrank Z-values at the six data-monitoring interim analyses were $Z(i)$ = 1.68, 2.24, 2.37, 2.30, 2.34, and 2.82, respectively.

Had the BHAT been designed to follow all the randomized patients until 400 events were observed — a maximum information trial — then the information fractions would have been 56/400 = 0.14, 77/400 = 0.19, 126/400 = 0.32, 177/400 = 0.44, 247/400 = 0.62, and 318/400 = 0.80. The corresponding monitoring boundary values for the six observations would have

Table 3. Interim analyses for the BHAT [10] trial using the alpha spending function $\alpha_1(t^*)$ with D = 400, T = 48

				Maximum information		Maximum duration	
Planned analysis	Calendar time (t months)	Total observed deaths (d)	Logrank Z	Information fraction (d/D)	Boundary value	Information fraction (t/T)	Boundary value
1	11	56	1.68	0.14	5.88	0.23	4.53
2	16	77	2.24	0.19	5.04	0.33	3.73
3	21	126	2.37	0.32	3.79	0.43	3.24
4	28	177	2.30	0.44	3.19	0.58	2.74
5	34	247	2.34	0.62	2.64	0.70	2.49
6	40	318	2.82	0.80	2.30	0.83	2.27
7	48	—	—	—	—	—	—

been 5.88, 5.04, 3.79, 3.19, 2.64, and 2.30. Again, this boundary would have been crossed at $t = 40$, or $t^* = 0.80$, with logrank statistic $Z = 2.82$.

However, since the BHAT was a maximum-duration trial of 48 months, we shall consider other ways to estimate the information fraction t^*. As indicated previously, one simple way to estimate t^* is by the fraction of calendar time. Let us reset calendar time $t = 0$ at June 1987, with the study ending in June 1982 when $t = 48$ (months). When data were monitored at time t between 0 and 48, we estimate t^* by $t/48$. This estimate may not be perfectly accurate, but it is simple to use. A slightly more accurate method for estimating t^* in a maximum-duration trial may be found in Lan and DeMets [40]. Note that the information fractions differ depending on which approach is used to design the trial.

In the original analysis, assuming equal increments from the O'Brien–Fleming [29] paper, the sixth of seven critical values was 2.20. If the test statistic had not exceeded the boundary value, it is possible that the DSMB might have called for another interim analysis at $t^* = 0.9$, for example. With this methodology, the exact boundary value could be computed.

The concept of information for the Wilcoxon test involves the joint distribution of censoring and survival time. When the mortality rate is low in a study, the information fraction of the logrank test gives a good approximation to the information fraction of the Wilcoxon test. In general, there is no simple interpretation of information for the Wilcoxon test. The interested reader should consult Lan, Rosenberger, and Lachin [57].

Repeated measures

Many trials consider outcomes other than a single mean value, an event, or time to failure. Trials may be designed in which a specific outcome (e.g., bone density, visual acuity) is measured repeatedly during the follow-up period. This design area, referred to as repeated measure design, has also been the subject of group sequential methods. Lee [58] provides an overview of this general class of methods. We shall focus on the most basic of repeated measures designs, namely, those that compare changes in a continuous response variable over time [59–64]. Wu and Lan [62] describe both linear and nonlinear mixed effects models.

Consider a trial in which, for each patient, several responses $y_1, y_2, \ldots, y_j$ are measured at successive follow-up times $x_1, x_2, \ldots, x_j$, and a least squares slope is computed to summarize the patient's response over time. A common model to analyze such a design would be a linear random effects model

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j \quad j = 1, 2, \ldots, J$$

for a specific patient, with β_0 being the intercept or constant, β_1 being the slope or change over time, and ε the deviation from the linear model. The ε_j's are assumed to be independent and normally distributed with mean 0 and variance σ_ε^2.The slope β_1 is assumed to be a random variable representing change over time, which varies from patient to patient. If β_1 varies across patients according to a normal distribution with mean B_1 and variance σ_β^2, then

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{J} (x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{J} (x_j - \bar{x})^2}$$

is an unbiased estimator of B_1 with variance

$$V(\hat{\beta}_1) = \sigma_\beta^2 \left\{ 1 + \frac{R}{\Sigma(x_j - \bar{x})^2} \right\},$$

where $R = \sigma_\varepsilon^2/\sigma_\beta^2$. The information from a single patient is

$$\left\{ 1 + \frac{R}{\Sigma(x_j + \bar{x})^2} \right\}^{-1}.$$

If we estimate a slope for each patient and take weighted averages across patients in the control group, then the estimated slope is

$$\bar{\beta}_C = \frac{\Sigma w_s \hat{\beta}_{1,S}}{\Sigma w_s}$$

where $w_s = V(\hat{\beta}_{1,S})^{-1}$. Expressions for the treatment group (T) are similar. In this case, the information fraction at the kth interim analysis would be

$$t_k^* = \{\mathrm{Var}(\hat{\beta}_C)^{-1} + \mathrm{Var}(\hat{\beta}_T)^{-1}\}/I$$

where I is the anticipated total information at the end if all observations are obtained. This figure is estimated in such a repeated-measures design. The test statistics Z_k is the standardized difference between the slope $\bar{\beta}_C$ and $\bar{\beta}_T$, computed across all patients and observations available at the point in time of the kth interim analysis:

$$Z_k = \frac{\bar{\beta}_{C(k)} - \bar{\beta}_{T(k)}}{\sqrt{V(\bar{\beta}_{C(k)} - \bar{\beta}_{T(k)})}}.$$

These test statistics are compared to the critical values obtained from the alpha spending function determined by the information fraction t_k^* for $k = 1, \ldots, K$.

An example of sequential monitoring under a simple linear mixed-effects model is provided by Lee and DeMets [60]. Bone density was repeatedly measured in a population of postmenopausal women to evaluate a calcium supplement treatment compared to a placebo control. Thirty-seven women were randomly assigned to receive calcium and thirty-seven to receive placebo. Bone density was measured on each woman 10 or 11 times over a five-year period [65]. Lee and DeMets [60] reanalyzed these data sequentially using a mixed-effects model fitting simple linear regression with a random slope coefficient for each woman. However, in their analysis, the number of measurements observed divided by the number expected was used as a surrogate for information fraction. Reboussin, Lan, and DeMets [66] repeated the analysis, but estimated the information fraction as described above. The expected total information was estimated by summing over the planned measurement times for each individual and then across individuals. Variances were estimated from these data, but would have had to be estimated from another source in the actual design of this trial. At each of five interim analyses, the observed information was computed using the variance estimates, and the information fraction was computed by dividing the observed the total expected. The test statistic computed at each interim analysis compares the two linear slope estimates of bone density decline. The spending function, $\alpha_1(t^*)$, was used but the corresponding boundary was truncated at a maximum value of 3.5. The results of these analyses are shown in table 4. As indicated, the test statistic exceeds the corresponding boundary for the spending function at $t^* = 0.77$ in the fourth interim analysis. Note that the information fractions are not equally spaced. If this monitoring procedure had been available, early termination of this trial might have been considered, although other factors might have argued for continuation. Note also the extreme value of the test statistic at the second analysis, which did not cross the monitoring boundary at that time and diminished in value at subsequent analyses.

Repeated confidence intervals

Above we have discussed the frequentist approach to group sequential monitoring from the hypothesis-testing point of view. An equally relevant approach is to calculate a confidence interval for the parameter of interest for treatment effect (e.g., difference in proportions, ratio of hazards) at each interim analysis [12,67–69]. As confidence intervals are computed, specific treatment differences of interest can be ruled in or out as possible values of the parameter. If all values of possible interest fall outside the

Table 4. Sequential analysis of repeated measurement of bone density study [65] (alpha = 0.05, $\alpha_1(t^*)$ spending function)

Interim analysis	Information observed	Information fraction (t^*)	Test statistic (Z)	Boundary value (Z_C)
1	0.24	0.01	0.38	3.50
2	2.61	0.11	3.14	3.50
3	8.83	0.37	2.33	3.50
4	18.48	0.77	2.49	2.31
5	24.15	1.00	2.19	2.02

confidence interval, the trial might be stopped with the conclusion that no difference of interest exists. However, if 0 should not be in the interval, we might stop and declare a treatment difference.

If the parameter representing treatment differences is denoted by θ, a nominal 95% confidence interval for a fixed sample design would have the general form $\hat{\theta} \pm 1.96\ \mathrm{SE}(\hat{\theta})$, where $\hat{\theta}$ is our estimate of θ and $\mathrm{SE}(\hat{\theta})$ is the standard error of the estimate $\hat{\theta}$. However, to use this nominal form repeatedly creates a similar type of problem as in the repeated testing approach: this nominal confidence interval will not have appropriate coverage. Thus, some adjustment must be made in order to compensate for the repeated application.

Jennison and Turnbull [67,68] have developed a repeated confidence interval (RCI) approach for the group sequential setting, modifying earlier sequential confidence interval approaches (see, for example, Robbins [23]). Formally, we want to construct a sequence of confidence intervals $[\underline{\theta}_k, \bar{\theta}_k]$ for θ such that

$$P_\theta\{\theta \subset [\underline{\theta}_k, \bar{\theta}_k] \text{ for all } k\} \geqslant 1 - \alpha.$$

That is, this sequence of intervals will collectively cover or include the unknown parameter with probability $1 - \alpha$.

Jennison and Turnbull [67,68] construct the repeated confidence intervals by inverting the group sequential test in which the critical value at the kth analysis $Z_\mathrm{C}(k)$ is determined by the alpha spending function. These RCIs are of the form

$$\bar{\theta}_k = \hat{\theta}_k - Z_\mathrm{C}(k)\mathrm{SE}(\tilde{\theta}_k)$$
$$\bar{\theta}_k = \hat{\theta}_k + Z_\mathrm{C}(k)\mathrm{SE}(\tilde{\theta}_k)$$

The coefficients $Z_\mathrm{C}(k)$ are the same critical values used for the repeated hypothesis testing. For example, for a 0.05 O'Brien–Fleming type boundary (as discussed above in the section on comparison of means), at the third ($k = 3$) interim analysis, the critical value was 2.63. Thus our RCI for θ would be $\hat{\theta}_3 \pm 2.63\ \mathrm{SE}(\hat{\theta}_3)$. The RCI and sequential test of H_0 will yield the same conclusions regarding the null hypothesis $H_0\!:\theta = 0$. However, RCIs provide more information about other possible values of the unknown parameter. For example, a DSMB may not want to terminate a trial unless they are sure that $\theta > \theta_0 > 0$; that is, the lower limit $\underline{\theta}_k > \theta_0 > 0$. This might occur if they judged that the treatment would have to have a difference much greater than 0 to compensate for coexisting toxicity.

This particular alpha spending approach to RCI has similar advantages as described for group sequential testing in that neither the timing nor the number of interim analyses needs to be specified in advance. The total expected information, I (e.g., a sample size of $2N$) must be determined for the design and used to calculate the information fraction for a specified alpha spending function. The RCIs are especially useful for equivalence trials [12,70–72] that are designed to test if two treatments have an effect within a specified acceptable difference and thus may be interchanged. That is, the treatments are 'equivalent' with respect to benefit, but one might be less expensive, less toxic, or less invasive.

The repeated confidence interval approach has been utilized in cancer and AIDS trials to establish equivalency [70–72]. For example, Fleming [72] describes the use of this concept by the Oncology Advisory Committee for the Food and Drug Administration (FDA). Federal regulations require cancer drugs to show an effect on survival, quality of life, or pain. Oncologists continue to seek new treatments or treatment combinations that may be 'equally effective' to the standard therapy but that offer an additional advantage such as being less toxic, less invasive, or more convenient to the patient. However, to establish a treatment as 'equally effective' requires setting a range of therapeutic equivalence — that is, a range of values for a relative risk (e.g., ratio of mortality in the new therapy compared to the standard) that oncologists would consider an even trade to gain the advantage of the new treatment. Often, the range of 0.8 to 1.2 for the relative risk is suggested as the definition of 'equally effective' or therapeutic indifference. Meier [24] discussed this idea, but did not adjust the confidence interval for repeated testing.

In this setting, we can imagine an equivalence trial in which RCIs are computed for each interim analysis. The trial would continue until the upper confidence limit for the relative risk was less than 1.2, meaning that we are reasonably (e.g., using 95% RCI for a 5% level alpha spending function) sure that the new treatment is not more than 20% worse than standard therapy. We might not have yet ruled out the possibility that the new therapy might even be superior; that is, the upper confidence limit is less than 1. However, if the RCI were contained in the rage (0.8, 1.2), we could terminate the trial, ruling out both a therapeutic advantage or disadvantage.

Sequential estimation

Once a trial has been completed, we would like to estimate the treatment effect. In the comparison of two means, the treatment difference is expressed by the difference between the mean responses (sometimes standardized by the standard deviation) from the two groups. In the survival setting, the hazard ratio is one way to indicate treatment difference, if it remains constant over time:

$$\text{hazard ratio} = (\text{hazard of group 1})(\text{hazard of group 2}).$$

For a fixed sample size or fixed information study, the observed treatment difference, which we will call the naive point estimate, at the end of the study is unbiased. The $(1-2\alpha)$ confidence intervals can also be constructed in the traditional way as

$$(\text{point estimate}) \pm z_\alpha(\text{standard deviation of point estimate}).$$

In general, construction of the point estimate or confidence intervals in the sequential setting is not so straightforward [73–83]. Naive estimates are biased after a sequentially designed trial has been completed, and appropriate adjustments for unbiased point estimates involve parameters whose values are typically unknown. Different proposals have been made to construct confidence intervals with correct coverage probability following a sequential test [73,74,77,79]. The authors of these proposals suggest different ways to order the sample space for sequential trials. The question is how to determine a treatment difference at one time point so that it is either more or less extreme than a difference at a second time point. In the Siegmund scheme [73], any result that exceeds the group sequential boundary at one time point is more extreme than any result that exceeds the boundary at any later time point. None of these methods is considered to be universally better than the others [79,82]. However, while the ordering suggested by Siegmund [73] and adopted by Tsiatis et al. [74] can break down, these cases are quite unusual [82]. Thus, we suggest using the method outlined in Tsiatis et al. [74].

Hughes and Pocock [83] pay particular attention to the fact that clinical trials that stop early are prone to exaggerate the magnitude of the treatment difference. They propose a Bayesian 'shrinkage' method, which uses a prior distribution to adjust the point and interval estimates. This approach requires a general agreement on the choice of prior distribution, however.

For sequentially designed trials, where there is a possibility of early termination, the amount of information obtained in the study may be less than that specified in the protocol. As a result, the power of a fixed design is greater than the power of a sequential design with the same maximum amount of information. Roughly speaking, there are two different types of strategy in sequential data monitoring. The aggressive one (the Pocock boundary is an example) puts more emphasis on early termination, and the conservative one (the O'Brien–Fleming boundary is an example) puts more emphasis on preserving power. The O'Brien–Fleming-type boundary is more commonly used for sequential data monitoring in many clinical trials. Such a conservative sequential plan is similar to a fixed-sample plan, and naive point and interval estimates are often adequate in practice. For aggressive sequential plans, one of the above-mentioned methods can be employed to reduce estimation bias. Sequential estimation is an important issue, and further research is still necessary.

Final remarks

In our experience over a variety of clinical trials, the alpha spending function implementation of group sequential interim monitoring has proven to be very helpful. It can be applied to most of the typical designs and analyses required in clinical trials and still has the necessary flexibility to meet the scientific and ethical needs of a data monitoring committee. Not all issues are totally resolved, however. One example is point estimation, described earlier. Another issue is what to do if the boundary values are crossed for the primary outcome, but the DSMB finds overwhelming reasons to continue [9,33,34,84]. From a statistical point of view, we can reject the null hypothesis no matter what Z-value is observed in the future. However, we find that most people feel uncomfortable with this approach and prefer to reject the null hypothesis only when the future Z-value exceeds a certain boundary. One suggestion [33,34] is to recapture all the previous alpha that has been spent and to redistribute it over the remainder of the trial.

In general, the alpha spending function approach is a generalization of previous versions of the group sequential approach, which provides not only control of the type error but also the flexibility required by the data monitoring process. If the alpha spending function with the total information or total duration is prespecified, the approach, while flexible for changes in the frequency and timing of analyses, is not subject to abuse. The alpha spending function approach has been used successfully in a wide variety of clinical trials that have often taken advantage of its inherent flexibility. While the decision to terminate or to continue a trial is a complex decision process, we recommend the alpha spending function as one factor in that process.

References

1. Friedman L, Furburg C, DeMets DL (1985). *Fundamentals of Clinical Trials*, 2nd edition. Littleton, MA: PSG.
2. Pocock SJ (1983). *Clinical Trials: A Practical Approach*. New York: Wiley.
3. Peto R, Pike MC, Armitage P, et al. (1976). Design and analysis of randomized clinical trials requiring prolonged observations of each patient. I. Introduction and design. *Br J Cancer* 34:585–612.
4. Heart Special Project Committee (1988). Organization, review and administration of cooperative studies (Greenberg Report): A report from the Heart Special Project Committee to the National Advisory Council, May 1967. *Controlled Clin Trials* 9:137–148.
5. Coronary Drug Project Research Group (1981). Practical aspects of decision making in clinical trials: The Coronary Drug Project as a case study. *Controlled Clin Trials* 1:363–376.
6. Beta-Blocker Heart Attack Trial Research Group (1982). A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. *JAMA* 247: 1707–1714.
7. Cardiac Arrhythmia Suppression Trial (CAST) Investigators (1989). Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N Engl J Med* 321(6):406–412.
8. DeMets DL (1990). Data monitoring and sequential analysis — An academic perspective. *J AIDS* 3 (Suppl 2):S124–S133.
9. DeMets DL (1984). Stopping guidelines vs. stopping rules: A practitioner's point of view. *Comm Stat (A)* 13(19):2395–2417.
10. DeMets DL, Hardy R, Friedman LM, Lan KKG (1984). Statistical aspects of early termination in the Beta-Blocker Heart Attack Trial. *Controlled Clin Trials* 5:362–372.
11. Pawitan Y, Hallstrom A (1990). Statistical interim monitoring of the cardiac arrhythmia suppression trial. *Stat Med* 9:1081–1090.

12. Fleming T, DeMets DL (1993). Monitoring of clinical trials: Issues and recommendations. *Controlled Clin Trials* 14:183–197.
13. Pocock SJ (1993). Statistical and ethical issues in monitoring clinical trials. *Stat Med* 12:1459–1469.
14. Pocock SJ (1992). When to stop a clinical trial. *Br Med J* 305:235–240.
15. Emerson SS, Fleming TR (1990). Interim analyses in clinical trials. *Oncology* 4:126–133.
16. Task Force of the Working Group on Arrhythmias of the European Society of Cardiology (in press). The early termination of clinical trials: Causes, consequences, and control — with special reference to trials in the field of arrhythmias and sudden death. *Eur Heart J.*
17. Lai TL (1984). Incorporating scientific, ethical and economic considerations into the design of clinical trials in pharmaceutical industry: A sequential approach. *Comm Stat (A)* 13: 2355–2368.
18. Wald A (1947). *Sequential Analysis*. New York: Wiley.
19. Bross I (1952). Sequential medical plans. *Biometrics* 8:188–205.
20. Anscombe FJ (1963). Sequential Medical Trials. *J Am Stat Assoc* 58:365–383.
21. Armitage P (1975). *Sequential Medical Trials*, 2nd edition. New York: John Wiley & Sons.
22. Armitage P, McPherson CK, Rowe BC (1969). Repeated significance tests on accumulating data. *J R Stat Soc A* 132:235–244.
23. Robbins H (1970). Statistical methods related to the law or iterated logarithm. *Ann Math Stat* 41:1397–1409.
24. Meier P (1975). Statistics and medical examination. *Biometrics* 31:511–529.
25. Canner PL (1977). Monitoring treatment differences in long-term clinical trials. *Biometrics* 33:603–615.
26. Whitehead J (1991). *The Design and Analysis of Sequential Clinical Trials*, 2nd edition. Chichester: Ellis Horwood.
27. Haybittle JL (1971). Repeated assessment of results in clinical trials of cancer treatment. *Br J Radiol* 44:793–797.
28. Pocock SJ (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* 64:191–199.
29. O'Brien PC, Fleming TR (1979). A multiple testing procedure for clinical trials. *Biometrics* 35:549–556.
30. Lan KKG, DeMets DL (1983). Discrete sequential boundaries for clinical trials. *Biometrika* 70:659–663.
31. DeMets DL (1987). Practical aspects in data monitoring: A brief review. *Stat Med* 6: 753–760.
32. Jennison C, Turnbull BW (1990). Statistical approaches to interim monitoring of medical trials: A review and commentary. *Stat Sci* 5:299–317.
33. DeMets DL, Lan KKG (in press). Interim analyses: The alpha spending function approach. *Stat Med.*
34. Lan KKG, DeMets DL, Halperin M (1984). More flexible sequential and non-sequential designs in long-term clinical trials. *Comm Stat (A)* 13(19):2339–2353.
35. Lan KKG, Zucker D (1993). Sequential monitoring of clinical trials: the role of information in Brownian motion. *Stat Med* 12:753–765.
36. Hwang IK, Shih WJ (1990). Group sequential designs using a family of type I error probability spending function. *Stat Med* 9:1439–1445.
37. Kim K, DeMets DL (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* 74:149–154.
38. DeMets DL, Ware JH (1982). Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika* 69:661–663.
39. Emerson SS, Fleming TR (1989). Symmetric group sequential test designs. *Biometrics* 45:905–932.
40. Lan KKG, DeMets DL (1989). Group sequential procedures: calendar versus information time. *Stat Med* 8:1191–1198.

41. Lan KKG, Reboussin DM, DeMets DL (in press). Information and information fractions for design and sequential monitoring of clinical trials. *Comm Stat (A)*.
42. Lan KK, Wittes J (in press). Data monitoring in complex clinical trials: which treatment is better? *J Stat Planning Inference*.
43. Li Z, Geller NL (1991). On the choice of times for date analysis in group sequential trials. *Biometrics* 47:745–750.
44. Lan KKG, DeMets DL (1989). Changing frequency of interim analyses in sequential monitoring. *Biometrics* 45:1017–1020.
45. Proschan MA, Follman DA, Waclawiw MA (1992). Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics* 48:1131–1143.
46. Kim K, DeMets DL (1992). Sample size determination for group sequential clinical trials with immediate response. *Stat Med* 11:1391–1399.
47. Fleming TR, Harrington DP (1991). *Counting Processes and Survival Analysis.* New York: Wiley.
48. Tarone R, Ware J (1978). On distribution free tests for equality of survival distributions. *Biometrika* 64:167–179.
49. Gail MH, DeMets DL, Slud EV (1992). Simulation studies on increments of the two-sample logrank score test for survival time data, with application to group sequential boundaries. In *Survival Analysis*, J Crowley, R Johnson (eds.), vol. 2. Hayward, CA: IMS Lecture Note Series.
50. Tsiatis AA (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *J Am Stat Assoc* 77:855–861.
51. Slud E, Wei LJ (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *J Am Stat Assoc* 77:862–868.
52. DeMets DL, Gail MH (1985). Use of logrank tests and group sequential methods at fixed calendar times. *Biometrics* 41:1039–1044.
53. Lan KKG, Lachin J (1990). Implementation of group sequential logrank tests in a maximum duration trial. *Biometrics* 46:759–770.
54. Sellke T, Siegmund D (1983). Sequential analysis of the proportional hazards model. *Biometrika* 70:315–326.
55. Kim K, Tsiatis AA (1990). Study duration for clinical trials with survival response and early stopping rule. *Biometrics* 46:81–92.
56. Kim K (1992). Study duration for group sequential clinical trials with censored survival data adjusting for stratification. *Stat Med* 11:1477–1488.
57. Lan KKG, Rosenberger WF, Lachin JM (1993). Use of spending functions for occasional or continuous monitoring of data in clinical trials. *Stat Med* 12:2214–2231.
58. Lee, JW (1994). Group sequential testing in clinical trials with multivariate observations: a review. *Stat Med* 13:101–111.
59. Laird NM, Ware JH (1983). Random effects models for longitudinal data. *Biometrics* 38:963–974.
60. Lee JW, DeMets DL (1991). Sequential comparison of change with repeated measurement data. *J Am Stat Assoc* 86:757–762.
61. Lee JW, DeMets DL (1992). Sequential rank tests with repeated measurements in clinical trials. *J Am Stat Assoc* 87:136–142.
62. Wu MC, Lan KKG (1992). Sequential monitoring for comparison of changes in a response variable in clinical trials. *Biometrics* 48:765–779.
63. Wei LJ, Su JQ, Lachin JM (1990). Interim analyses with repeated measurements in a sequential clinical trial. *Biometrika* 77(2):359–364.
64. Su JQ, Lachin J (1992). Group sequential distribution-free methods for the analysis of multivariate observations. *Biometrics* 48:1033–1042.
65. Smith E, Sempos CT, Smith PE, Gilligan C (1989). Calcium supplementation and bone loss in middle-aged women. *Am J Clin* 50:833–842.
66. Reboussin D, Lan KKG, DeMets DL (1992). Group sequential testing of longitudinal data. Technical Report #72, Department of Biostatistics, University of Wisconsin, Madison, WI.

67. Jennison C, Turnbull BW (1984). Repeated confidence intervals for group sequential trials. *Controlled Clin Trials* 5:33–45.
68. Jennison C, Turnbull BW (1989). Interim analyses: The repeated confidence interval approach. *J R Stat Soc B* 51:305–361.
69. DeMets DL, Lan KKG (1989). Discussion of: Interim analyses: The repeated confidence interval approach by C. Jennison and B.W. Turnbull. *J R Stat Soc B* 51:362.
70. Fleming TR, Watelet LF (1989). Approaches to monitoring clinical trial. *J Natl Cancer Inst* 81(3):188–193.
71. Fleming TR (1990). Evaluation of active control trials in AIDS. *J AIDS* 3 (Suppl):S82–S87.
72. Fleming TR (1978). Treatment evaluation in active control studies. *Cancer Treat Rep* 17(11):1061–1065.
73. Siegmund D (1978). Estimation following sequential tests. *Biometrika* 65:341–349.
74. Tsiatis AA, Rosner GL, Mehta CR (1984). Exact confidence intervals following a group sequential test. *Biometrics* 40:797–803.
75. Rosner GL, Tsiatis AA (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika* 75:723–729.
76. Kim K (1989). Point estimation following group sequential tests. *Biometrics* 45:613–617.
77. Kim K, DeMets DL (1987). Confidence intervals following group sequential tests in clinical trials. *Biometrics* 4:857–864.
78. Whitehead J (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika* 73:573–581.
79. Emerson SS, Fleming TR (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika* 77:875–892.
80. Pocock SJ, Hughes MD (1989). Practical problems in interim analyses, with particular regard to estimation. *Controlled Clin Trials* 10:2094–2215.
81. Chang MN, O'Brien PC (1986). Confidence intervals following group sequential tests. *Controlled Clin Trials* 7:18–26.
82. Whitehead J, Facey KM (1991). Analysis after a sequential trial: a comparison of orderings of the sample space. Presented at the Joint Society for Clinical Trials/International Society for Clinical Biostatistics, Brussels.
83. Hughes MD, Pocock SJ (1988). Stopping rules and estimation problems in clinical trials. Stat Med 7:1231–1241.
84. Whitehead J (1992). Overrunning and underrunning in sequential trials. *Controlled Clin Trials* 13:106–121.

2. Issues in the design and analysis of AIDS clinical trials

Dennis O. Dixon and Jeffrey M. Albert

Introduction

The AIDS epidemic has provoked a massive response from the worldwide scientific community. To address the need for rapid evaluation of new treatments for the primary viral infection and related opportunistic infections, malignancies, and other illnesses, the Federal Government established the largest publicly-sponsored program of clinical trials ever undertaken. As intended, these resources made it possible to enlist many gifted academic and government scientists in the effort, including biostatisticians and other clinical trialists. There have been advances in connection with several aspects of clinical trial design and analysis, and the body of this chapter highlights a few of these. Before beginning, however, we consider the question of why it is necessary, or at least useful, to focus on advances in methodology for AIDS clinical trials.

The short answer is that it is not necessary at all; trials methodology applies regardless of disease. On the other hand, issues inevitably rise to prominence earlier in one disease area than another, according to sometimes surprising external considerations.

An indication of the way in which innovations in methodology developed in HIV-related clinical research affects research on other diseases is the following comment by Dr. Anthony S. Fauci, then Director of the National Institutes of Health Office of AIDS Research:

> . . . new approaches for conducting clinical trials have been developed, including community-based trials, which capture the expertise of community physicians. Similarly, AIDS clinical trials have demonstrated the importance of providing ancillary services to recruit and ensure the continued participation of underserved populations such as women, children, minorities, and injecting drug users (IDUs) in clinical trials for other diseases. . . . Community advisory boards assist clinical trials sites in ensuring close cooperation with community constituency groups.

Peter F. Thall (ed.), RECENT ADVANCES IN CLINICAL TRIAL DESIGN AND ANALYSIS.
Copyright © 1995 Kluwer Academic Publishers, Boston. All rights reserved. ISBN 978-1-4613-5830-5.

> Other innovations brought about by AIDS-related research include the identification of surrogate markers that reliably provide valid information for determining the outcome of clinical studies. . . .
>
> In all of these efforts an integral role is played by AIDS activists and advocates for HIV-infected individuals. AIDS-community representatives serve on all AIDS-related advisory committees and on protocol design teams of clinical trials in an unprecedented collaboration between researchers and those affected and infected by HIV ([1], pp. 5–6).

In this chapter we review work in several areas. We discuss in some detail research on the identification and use of surrogate markers, methods related to the study of quality of life, and methods for dealing with noncompliance, since these areas are so important in HIV/AIDS trials. Next, we issue an appeal for fresh thinking and new approaches with regard to the effort to improve the efficiency of clinical trials, not from the point of view of experimental design, but by improving the process of obtaining accurate and complete data to analyze. Finally, we comment more briefly on several other topics that have received significant attention.

Other recent reviews and commentaries include those by Byar et al. [2], Green et al. [3], Ellenberg et al. [4], Ellenberg, Finkelstein, and Schoenfeld [5], Dixon et al. [6], and Ellenberg and Dixon [7].

Surrogate endpoints

The efficacy of a potential HIV/AIDS therapy would ideally be assessed through clinical endpoints such as survival or the occurrence of an opportunistic infection (OI). However, the use of clinical endpoints, except for populations in an advanced disease stage, requires a very lengthy follow-up period and/or large sample sizes. Furthermore, dropouts and treatment changes or withdrawals often occur before death or progression to AIDS, obscuring the assessment of treatment efficacy. In light of the urgent need to rapidly identify and evaluate potential HIV disease therapies, as well as to utilize limited resources optimally, more quickly attainable 'surrogate endpoints' have been sought.

In a broad sense, a surrogate endpoint may be any variable chosen to replace a desired endpoint, usually for expediency or cost reduction. Thus, for example, quality of life or occurrence of opportunistic infection may be considered as surrogates for death due to AIDS. Of course, such measures of well-being can be justified as primary endpoints in their own right. On the other hand, nonclinical (usually laboratory) measurements are not of direct concern to the patient and thus provide the more contentious (and hence prototypical) class of surrogate endpoints.

Laboratory variables used in the study of AIDS include virological markers, such as PCR assay for HIV DNA, plasma viremia, and level of HIV p24 core antigen; and immunological markers, such as CD4+ lymphocyte counts, serum β_2-microglobulin and serum neopterin levels. CD4 counts (sometimes relative to CD8+ or total lymphocyte numbers) have received the most attention as a potential surrogate endpoint. Already in the AIDS Clinical Trials Group (ACTG), CD4 counts have been used in the definition of primary endpoints in phase III clinical trials. Indeed, the recently updated definition of AIDS by the Centers for Disease Control and Prevention (CDC) allows for determination to be based in part on the CD4 count.

Much discussion in the AIDS literature regarding surrogate endpoints has been given to defining criteria for determining the desirability and/or acceptability of their use. A broad perspective is given by Amato and Lagakos [8], who provide a list of considerations in choosing an endpoint for a clinical trial. Briefly, these include 'relevance' (how well the endpoint indicates actual treatment benefit), time needed to evaluate the endpoint, verifiability or objectivity, variability, and implementability. Laboratory measurements such as CD4 counts, in addition to being relatively quickly responsive to drug therapy, have the advantage of being objectively obtainable, as opposed to 'softer' endpoints such as quality of life or even the presumptive diagnosis of AIDS. Among the disadvantages of laboratory markers are their requirement of expensive, 'high tech', equipment (this is particularly the case for virological markers and especially problematic for community-centered clinical trials); high variability; and questionable relevance to the patient's well-being.

The question of relevance is perhaps the most critical when considering a laboratory marker as a surrogate endpoint (or 'surrogate marker') and has been the central focus of marker 'validation.' Another way of expressing this requirement is that a surrogate marker should be a 'measure central to the pathogenesis of disease' and should furthermore have a causal connection to clinical outcome [9]. Meeting this criterion would imply, as desired, that the marker would predict the clinical benefit of a wide range of drugs. Merigan (see [10]) proposed three quantitative criteria that would support such a causal connection: (1) the marker should be monotonically affected by drug dose; (2) it should correlate with disease progression in untreated patients; and (3) changes due to treatment should correlate with clinical effect. In practice, as suggested by Merigan's criteria, the validation of a surrogate marker is tied to a given drug, or at best, a *class* of drugs. Evidence of a marker's causal connection to outcome does not preclude the possibility that other drugs might affect the clinical endpoint via a mechanism that does not involve the marker.

For some purposes, an endpoint that fulfills some but not all of Merigan's criteria may be of interest. A variable satisfying condition (2) would be considered a prognostic factor or 'type 0' marker [11] and may be useful in monitoring the course of disease. Conditions (1) and (2) may be sufficient to define an 'activity' or 'type I' marker [11] that may be deemed an adequate endpoint for a phase I or phase II clinical trial. The presence of all three conditions establishes a surrogate endpoint in the present sense (also referred to as an 'efficacy' or 'type II' marker [11]). In particular, as demonstrated in a simulation study by Machado et al. [12], a variable that is highly prognostic for a clinical endpoint, but does not satisfy condition 3, may in fact be misleading with regard to treatment effect on the clinical endpoint.

A criterion intended to validate a surrogate endpoint was delineated and operationalized by Prentice [13]. Prentice's criterion (roughly indicated by condition 3 above) essentially requires that the clinical endpoint be independent of treatment given the surrogate outcome. This condition (if, in addition, the marker is prognostic for the clinical outcome) implies that the surrogate produces a valid test of the null hypothesis of no association between treatment and clinical outcome.

Several investigators have examined CD4 level as a surrogate marker in AIDS clinical trials employing this criterion. In an analysis of data from ACTG Protocol 019, Choi et al. [14] found, using a Cox regression model, that only a small proportion of the effect of ZDV on the progression to AIDS of asymptomatic HIV-infected patients was explained by its effect on CD4 counts. An estimated 'explained' proportion was obtained as the proportional reduction in the regression coefficient for the treatment assignment after controlling for CD4 levels. These proportions were 0% for current CD4 counts and 37% using net CD4 counts (percentage of CD4+ lymphocytes among all leukocytes). Choi et al. concluded that CD4 levels provide an 'incomplete' surrogate marker for disease progression in the context of ZDV treatment of asymptomatic HIV-infected patients.

Lin et al. [15] also utilized a Cox regression model with CD4 counts as a time-varying covariate, but 'censored' the CD4 count if it was not taken within a sufficiently recent time window (specified according to the pattern of available data). They found that CD4 counts fail (by Prentice's criterion) as a surrogate endpoint for progression to OI or AIDS in advanced disease populations. They also note that the same result is obtained using *change* in CD4 count relative to baseline rather than current CD4 count. Similar conclusions, using a model accounting for CD4 measurement error, were obtained by DeGruttola et al. [16].

While the Prentice criterion may be regarded as too stringent and unrealistic, it leads, as we have seen, to a useful measure of the adequacy of a marker as a surrogate endpoint, namely, the percentage of treatment effect on clinical outcome explained by the treatment effect on the marker. One limitation of this perspective based on hypothesis testing is that it does not address the use of a marker to provide information about the *magnitude* of the effect of treatment on the clinical endpoint.

The inadequacy of the most promising surrogate marker in the AIDS arena (and the likely failure of future markers) in meeting the ideal of the Prentice criterion reinforces doubts about this usage of laboratory variables.

Of course, this concern is greatest in phase III trials that are intended to yield definitive conclusions about treatment benefit, though it is also relevant to earlier phase trials. What is more, past experience in AIDS and other disease areas has induced a wariness about the use of surrogate endpoints. As continued study of ZDV and other anti-HIV treatments has dampened the initial enthusiasm for these drugs, there is question as to how much we may have been misled due to the widespread and early reliance on surrogate endpoints, and on the CD4 levels in particular (see, e.g., [17]).

With this increasingly cautious attitude about surrogate endpoints has come an increased interest in methods that use prognostic laboratory observations other than as outright substitutes for clinical endpoints. In one approach, associated with efforts to evaluate quality of life, marker information is combined with information on clinical endpoints. As a simple example, a treatment failure may be considered to occur if either the primary clinical endpoint occurs or the marker variable meets a specified threshold. This combined endpoint gives equal weight to the clinical and laboratory variables. More generally, weights may be assigned to reflect the relative importance of the respective outcomes.

An alternative is to use laboratory markers as *auxiliary* information to strengthen an analysis based on the desired clinical endpoint. Kosorok and Fleming [18] consider a situation in which patients are randomized both to treatment and to either limited or extended follow-up. Patients in the former group are either followed for a shorter period of time or are followed for a secondary failure endpoint (a generally earlier-occurring surrogate to the primary endpoint). In such a design, a cost reduction is achieved by the shorter follow-up of a subgroup of patients, but this advantage must be weighed against the loss of information about the primary endpoint. Kosorok and Fleming present a nonparametric test statistic (representing a linear combination of statistics based on the two endpoints) that utilizes information on the auxiliary variable without introducing bias. From their simulation study, they find that correlations greater than 0.7 are needed to gain substantially improved power from the use of the auxiliary variable.

Fleming et al. [19] (see also [20,21]) developed other approaches that allow more flexibility in the definition of the auxiliary variable; for example, CD4 counts over time may be utilized rather than time to a defined failure endpoint, as in the Kosorok and Fleming method.

An approach by Finkelstein and Schoenfeld [22] incorporates information on an auxiliary failure-time endpoint in a modified Kaplan–Meier estimator and use this estimator to obtain a modified logrank statistic to test for treatment effect. This approach does not rely upon randomization to different lengths of follow-up. Simulation results indicate at best modest gains, and in some situations *losses*, in efficiency relative to standard methods.

Thus, while gains in power may be realized, they are largely confined to special cases in which the auxiliary endpoint occurs much earlier than and is highly predictive of the primary endpoint. Kosorok and Fleming suggest that the use of multiple secondary endpoints (readily accommodated by their methodology) might further improve statistical power, though one must be attentive to possible multicollinearity.

As a final consideration, the use of surrogate markers should reflect the regulatory and administrative processes through which treatment assessment takes place. For example, surrogate endpoints may be used to provide tentative decisions regarding drug approval or treatment recommendation, under the condition that follow-up be continued to allow validation based on long-term clinical outcomes. Implications of the use of surrogate markers in interim monitoring also require further study.

Quality of life methodology

Increase in the length of life, the health-related quality of life, or both is the universal aim of therapeutic intervention. Sometimes, however, a treatment may prolong survival but diminish quality of life through toxicity, or may even lack survival benefit but postpone symptoms of advanced disease at a cost of short-term toxicity. HIV infection is invariably fatal with presently available treatments, but lifespans are measured in years. It is thus especially important to address impact of treatments upon quality of life, and there is a great deal of interest at present in methodology for evaluating treatments for HIV infections from this perspective.

Quality of life improvement, suitably defined, might serve as a surrogate marker for subsequent clinical outcome in the sense of the last section, or as a therapeutic objective in its own right. In the latter case, it may then be necessary to synthesize results of several types of comparisons of treatments in some formal way, including multivariate analysis.

'Quality of life' has been a difficult concept to measure, or even to define. For some purposes, a simple overall indicator of performance status, or the extent to which a person can carry out normal activities of daily living, may be sufficient. In other cases, it may be of interest to develop multidimensional, context-specific, self-assessment scales in order to adequately express quality of life numerically. Guyatt et al. [23] propose a taxonomy of measures and discuss the advantages and disadvantages of generic instruments, including both health profiles and utility measurements, and situation-specific choices.

Among those advocating the use of multidimensional questionnaires, Fitzpatrick et al. [24] assert, 'In clinical trials many scientific questions cannot be answered properly without adequate measurement of quality of life,' and list several requirements of such measurements. These are reliability, validity, sensitivity to change, appropriateness, and practicality. Elsewhere [25], the same group recommended use of a validated standard measuring instrument supplemented with customized additions relevant to a particular situation.

Williams and Rabkin [26] evaluated a Quality of Life Index (QLI) that Spitzer et al. [27] had devised for cancer patients, in a cohort of 50 gay men, 29 of whom were HIV positive. QLI summarizes self-assessments of levels of activity, daily living, health, support, and outlook. By comparing it with other standard measures of functioning, Williams and Rabkin concluded that QLI succeeds in capturing relevant information important to this population, thus establishing validity and appropriateness. They had not yet examined sensitivity to changes attributable to treatment for the HIV infection.

Wu et al. [28] proposed a 30-item questionnaire addressing 10 aspects of health developed by adding items to the Medical Outcomes Study short-form General Health Survey. They recognized that many health status instruments are too lengthy for use in clinical trials in HIV disease, but that no single observation could reliably capture health status.

In order to evaluate the proposed questionnaire, Wu et al. [28] studied 73 volunteers with asymptomatic HIV infection and 44 volunteers with early AIDS-related complex (ARC), all of whom had enrolled in controlled clinical trials carried out by the AIDS Clinical Trials Group (ACTG). To assess validity they compared responses between groups on the 10 scales using rank-sum tests. Asymptomatic patients had significantly better overall health, better physical and role function, less pain, better cognitive functioning, and better quality of life than patients with early ARC; marginally significantly better energy/less fatigue; and similar social functioning, mental health, health distress, and health transition.

Each of the various health status scales, as well as many other indicators of physical condition, provide a basis for comparing groups receiving different treatments. While it is possible to make such comparisons one at a time, the ultimate decision made by the individual, together with his or her physician, requires a balancing of advantages and disadvantages of the treatment alternatives.

Glasziou, Simes, and Gelber [29] discuss methods for synthesizing results of comparing groups on the basis of several outcome measures. They consider situations in which quantitative differences can be expressed on a single scale and assume that one can specify weights corresponding to the relative utilities of advantages of the different outcomes. An important special case occurs when health-related quality of life data can be reduced to a small number of clearly ordered health states, and individuals are observed to spend varying amounts of time in the different states.

Gelber, Gelman, and Goldhirsch [30] discuss the statistical properties of quality-adjusted time without disease symptoms or treatment toxicity (q-TWiST), i.e., overall survival discounted for periods of time spent with reduced quality of life. Two large clinical trials of antiretroviral treatments for HIV have recently been reanalyzed using q-TWiST.

ACTG 016 provided evidence that persons with mildly symptomatic HIV infections remained progression-free longer if they received zidovudine

(1200 mg/d) than if they received placebo [31]. On the other hand, they spent longer times with severe symptomatic adverse events. Gelber et al. [32] found that over the 18-month observation period of the trial, 'treatment provided more q-TWiST than placebo if the quality of life after HIV disease progression was assumed to be 10% to 20% worse than the quality of life after a severe symptomatic adverse event.'

A trial in persons with asymptomatic HIV infections and less than 500 CD4-positive cells per mm^3 had also demonstrated a delay in clinical progression associated with zidovudine (500 mg/d) [33]. In this case, however, Lenderking et al. [34] have concluded, on the basis of q-TWiST analysis, that 'a reduction in the quality of life due to severe side effects of therapy approximately equals the increase in the quality of life associated with a delay in the progression of HIV disease.'

Noncompliance

Patient noncompliance is a heterogeneous and multidimensional phenomenon. It may involve any of the multitude of components of a typical treatment regimen: taking too little (or too much) of the prescribed medication, failure to adhere to the assigned schedule, taking prohibited medications, or missing scheduled clinic visits. The problem of patient noncompliance with study regimen has been recognized in many disease areas (see, e.g., [35]). It is of particular concern in HIV/AIDS, where an atmosphere of rapidly alternating hopes and disappointments adds to the difficulty of keeping patients on a fixed, long-term treatment schedule. Inspired in part by AIDS activists, potential study participants in AIDS and other disease areas are increasingly active and assertive in opposing study designs or violating protocol rules not deemed to be in their best interest.

Although the label *noncompliant* is generally perceived as pejorative, the broader concern is with any departure from intended treatment, whether capricious or justified and intelligent. For example, for the current antiretroviral drugs, high levels of toxicity often require early treatment withdrawal. Furthermore, a treatment arm may be stopped early or modified due to interim analysis results or other information about treatments in the study. This occurred in the Concorde Trial [17] in which patients originally assigned to delayed zidovudine were offered the drug early as a result of positive findings in a separate trial.

Noncompliance, in the general sense of departure from intended treatment, has serious implications for the analysis and interpretation of clinical trials results. One consequence is that the standard biologic interpretation of treatment effect may no longer be appropriate. The presence of treatment noncompliance implies that the test comparing treatment groups *as randomized* may not adequately reflect the effect of a therapy taken as intended ('efficacy').

The intent-to-treat or as randomized analysis provides an assessment of treatment *effectiveness*. Treatment effectiveness, so defined, is determined by compliance on treatment, as well as its biologic effect. The issue of noncompliance with regard to the intent-to-treat analysis is one of statistical power. Often, power is determined for the detection of specified treatment efficacy. Noncompliance generally erodes this power; thus, larger sample sizes are required to detect the same level of efficacy. Equivalently, desired power is obtained by calculating the sample size to detect a given *effectiveness* rather than efficacy. Methods for adjusting sample sizes in the presence of noncompliance are provided, for example, by Schork and Remington [36] and Lachin and Foulkes [37].

The estimation of efficacy is not so straightforward and is generally approached by some sort of *as-treated* analysis, which uses information on the amount or pattern of treatment actually received. Statisticians, in particular, have warned of the likely bias in an analysis that compares treatments on the basis of a postrandomization variable such as compliance [38–40]. As in one classic example [41], it is often the case that compliers on placebo have better outcomes on average than placebo noncompliers. It is generally advisable to present an intent-to-treat analysis, even if treatment efficacy is of primary interest. An as-treated approach may be considered in addition as a secondary, exploratory analysis.

Apart from simple descriptive statistics, few as-treated analyses of AIDS clinical trials data have been published. However, we review here some recently proposed methods that provide the most promising directions for AIDS and perhaps other clinical trials.

To begin, we reflect on the problems of a naive analysis (such as a simple comparison of compliers on active treatment and placebo) that a more sophisticated as-treated analysis would seek to overcome. First, there may be heterogeneity of treatment effects so that the benefit of active treatment for people who tend to comply may be different from the benefit that *would have* been obtained by the noncompliers had they taken the full treatment. Secondly, proportions of compliers (more generally, the compliance distribution) may be different for different treatments (or placebo). Often this will undermine the comparability of compliers on different treatments. For example, due to drug toxicity, there may be a reduced proportion of compliers on an active drug relative to placebo. Since compliers on the drug are those less vulnerable to or more tolerant of side effects, they are apt to be healthier as a group than compliers on placebo. Finally, there may be different relationships between compliance and a confounding variable (health at baseline, say) for the different treatments. As an example, compliers on a given active treatment may be healthier than compliers on placebo (even if there are the same number in both groups), possibly because compliers on active treatment are better able to tolerate side effects, while compliers on placebo may be the sicker patients who are more motivated to adhere to a treatment regimen.

The method of Efron and Feldman [42] deals with the first two of these problems. They address the situation involving quantitative response and compliance variables that are obtained for patients randomized both to treatment and placebo. The second problem ('marginal noncomparability') is handled by transforming compliance on placebo so that it takes the value corresponding to the same *percentile* of compliance on treatment. Thus, an individual at, say, the 75th percentile for compliance on placebo is given the score at the 75th percentile of treatment compliance. This device produces the same distribution of compliance for both treatment groups. The Efron and Feldman approach then essentially compares average treatment responses based on individuals at the same percentile of compliance for the different randomization groups.

Generally, this transformation will not assure comparability. In particular, the resulting estimators (of conditional expected treatment effect for a given drug dose or compliance level) are prone to bias in the presence of unknown confounding factors (the third problem above). The possible bias was illustrated in a hypothetical example by Mark and Robins [43] and was related to the degree of departure from assumptions in a simulation study by Albert and DeMets [44].

Efron and Feldman attempt to deal with possible heterogeneity by utilizing a model that describes the *causal* treatment effect for an individual as a linear function of placebo response and possibly baseline covariates. Consequently, estimates of *conditional* expected treatment effect for patients at a given compliance percentile are used to draw inference on the expected treatment effect for the whole population (of primary interest but not directly estimable). The difficulty arises from the fact that drug and placebo responses are not both observed for a given individual, as would be required to directly measure a causal effect. Thus, the available data cannot entirely distinguish between heterogeneity (differences in patients who comply at different levels) and the effect of varying drug dose, so the (realistic) causal model will be nonidentifiable.

An alternative approach was proposed by Mark and Robins [43] (see also [45]) for the context of a failure-time endpoint. This method also involves a causal model — in this case, one describing the failure time for an individual as a function of his or her (possibly unobserved) compliance level. However, this approach avoids comparisons based on potentially noncomparable subgroups. Instead, it relies on the comparability induced by randomization to assess the causal impact of treatment (according to its measured level) on failure time. Mark and Robins provide a logrank-type test for efficacy that maintains the nominal false-positive rate under the null hypothesis of no treatment effect (however, this is in the strong sense of no effect on any individual). Thus, a valid statistical test of efficacy is available; however, the estimate of efficacy must be viewed with caution, given the speculative nature of the model on which it is based.

Other approaches to the assessment of efficacy for a failure-time endpoint are discussed by Peduzzi et al. [38]. These methods also construct logrank statistics, but unlike the Mark and Robins approach essentially alter individual treatment categorizations over time depending on the actual (observed) treatment received. The most important of these methods are the censoring method, which censors an individual at the time of treatment change or withdrawal, and the transition method, in which the comparison at a given point in time is based on the treatments individuals are observed to be on at that time point. As with the Efron and Feldman method, these approaches are prone to bias in the presence of confounding variables. Whereas the Efron and Feldman method summarizes information to obtain a single compliance score, the censoring and transition methods make use of the pattern of compliance over time. Consequently, the sort of transformation used by Efron and Feldman to improve comparability is not directly applicable to the latter methods.

In an approach that attempts to avoid the biases of most estimators of efficacy, Lagakos et al. [46] proposed a method for improving the power of an intent-to-treat type analysis that takes into account the expected pattern of noncompliance (i.e., the probability of withdrawing from treatment as a function of time). They developed a weighted logrank statistic that is weighted (in a manner to optimize power) according to this expected non-compliance. This test retains the nominal false-positive rate, and simulation results demonstrated its potential under certain circumstances for decreasing required sample sizes by up to 30%. The statistic does not use contemporaneous compliance data, but relies on relevant compliance information obtained a priori. The use of such a method thus motivates the routine collection of compliance information. While the method preserves the unbiasedness of the intent-to-treat analysis, it is not really directed at treatment effectiveness. Rather, it attempts to weight more heavily those periods of time when effectiveness is less dampened by noncompliance and thus to be more reflective of treatment efficacy. The method does have its potential for abuse and should accordingly be accompanied by the hypothesized patterns of noncompliance and efficacy.

An obvious remaining issue is the problem of accurately capturing compliance or treatment actually received. A number of methods have been proposed, and most have already been used in AIDS clinical trials. These methods include pill counts; electronic medication event monitoring (e.g., MEMS caps); pharmacological testing for drug, drug metabolites, or markers; records of missed appointments; and patient or physician reports. Such measurement tools have been extensively studied and reviewed (see, e.g., [47]).

Several novel approaches to constructing compliance variables and estimating compliance have been proposed for AIDS clinical trials. Lim [48] developed an estimate of the overall compliance rate using serum drug levels that corrects for false positivity and false negativity in drug determinations. This estimate makes use of information about error rates of the assay and the pharmacokinetics of the drug.

Richardson et al. [49], proposed a method using uric acid as a marker for ddI serum levels. Uric acid levels are elevated by ddI and would potentially provide a less expensive measurement of ddI compliance than ddI serum levels. In discriminant analyses intended to validate this use of uric acid (against ddI serum levels), these investigators showed that a linear function utilizing uric acid levels correctly identified as compliers or noncompliers 84% and 75%, respectively, of the patients in two test groups.

The assessment of compliance may be greatly enhanced though the use of multiple measuring tools. In a substudy of ACTG 175, a multicenter clinical trial involving asymptomatic HIV-infected patients, compliance on ZDV and ddI is being investigated using both MEMS caps and serum concentration determinations. Serum concentrations are measured at one randomly chosen site each week. MEMS caps are used at two specially selected sites. Using prior knowledge of the pharmacokinetics of ZDV and ddI as well as methods of population pharmacokinetics, overall estimates of compliance can be obtained. In addition, the investigators proposed to use Bayesian techniques to estimate a compliance profile for each patient. Subsequently, as-treated analyses may be performed to relate estimated drug exposure to outcome.

As we have suggested, efforts to monitor compliance may be worthwhile whether the primary question is one of efficacy or effectiveness. While compliance data are used directly in inference regarding efficacy, they provide information for the design of future trials intended to assess effectiveness. Furthermore, compliance is an interesting endpoint in its own right and may provide important information about the willingness of patients to carry out a prescribed therapeutic regimen.

Thoughts on data quality in clinical trials

Especially in multicenter clinical trials systems, large amounts of resources are devoted to editing and correcting data in preparation for analysis. These efforts are, in our view, somewhat misguided; many of the most critical errors found at this stage cannot be corrected. In what follows, we argue for redirection of most resources to improve data quality early — that is, to clarify protocols, simplify forms, and, most importantly, improve training. Clinic staff need procedures to help them find errors while there is still time for correction and to permit them to reduce error rates generally. We must devise methods to document the high quality of data leaving the research units in a way that is convincing, so that inspection on a large scale is unnecessary.

Responsibility for data quality is shared by those designing studies, those conducting studies, coordinating centers, sponsors, and arbiters (such as journal editors and the U.S. Food and Drug Administration). Each takes certain steps to control, assure, or enhance quality. It must be clear that the sharing of responsibility risks the abdication of responsibility.

There is very little attention paid to study of the effectiveness of procedures for quality control. What evidence is available that protocols and forms are well constructed, that research centers are following protocols, that coordinating centers are detecting problems early with good effect, that data entry errors are avoided, or that FDA approval to market a new drug for a particular indication is based on data of high quality?

Most multicenter trials, including almost all sponsored by NIH, involve a coordinating center, which conducts centralized as well as on-site reviews of data at some time after submission. Reviews may consist of any combination of computerized edit checks, review of cases by data managers, and item-by-item scrutiny of research records in comparison with source documents. The checking is inevitably delayed from the actual clinic visit, laboratory testing, and completion and submission of forms.

The pharmaceutical industry sponsors many trials and takes primary responsibility for data quality for most of them. The usual arrangement is for regionally based company employees (often with the title Clinical Research Associate) to visit each research unit every 1 to 4 weeks to check all data generated since the last visit. To the extent possible, corrections are made before data ever enter the research data base. The CRAs in effect supplement the research unit staff. The company may also conduct separate quality assurance reviews by staff from an administratively independent part of the company.

The industry model can be thought of as an attempt to make inspections more complete and, especially, more timely.

If results of a study form a critical part of an application for FDA approval to promote a treatment for a specified condition, FDA staff will ordinarily review records at one or more participating research units in order to verify the quality of data.

In order to evaluate the usefulness and the limitations of the systems described above, it is first necessary to understand the kinds of problems that occur with clinical trial data, how they can be detected, and what can be done about them if detected.

In many ways a clinical trial is meant to be a scientific experiment with as many features as possible in common with a laboratory experiment. The protocol is the experimental plan, complete with detailed specification of the types of volunteers who will participate, the treatment(s) to be given, and the types of evaluations to be performed. Interpretation of results is more or less straightforward according to the degree to which investigators and volunteers follow the protocol.

Potential departures from the protocol include enrollment of volunteers not meeting eligibility criteria, failure to administer treatments according to instructions (including those for managing adverse experiences), failure to carry out specified follow-up visits, failure to conduct specified clinical or laboratory tests, failure to abide by specified decision algorithms (e.g., for assessing response to or failure of treatment), and failure to report observations completely, accurately, and promptly. Ultimately, the investigators need to be able to assert that they conducted a true evaluation of the specified experimental treatment(s).

When departures do occur, as they inevitably will, it is helpful to know why. Occasionally, departures occur because the protocol fails to make provision for some aspect of disease presentation or some reaction to the interventions under study. In such cases, best clinical judgment takes precedence.

More often, however, the problem is that clinic staff simply do not clearly understand what the protocol intends. For example, a laboratory test result might be missed because the individual responsible for scheduling clinic visits overlooked the requirement to order a particular test at the time of a particular visit. Or the protocol might specify that, after an initial period of time, follow-up visits change from every two weeks to every three weeks. If the volunteer is asked to return in two weeks, by mistake, it might be too great an imposition to ask him or her to come back one week later in order to be tested 'on time.'

Of course, it might also happen that the test was done but that the result was not recorded on a study form.

Data checking is the effort to discover errors of all kinds in the research data. Other than filling out report cards, however, there is effectively no opportunity to correct most errors, even if they are found. Returning to the earlier example, a test not done when specified cannot be made up.

The retrospective and remote checking of data does not necessarily eliminate all or even most errors, as illustrated in a fascinating study reported by Pritzker [50] (and recounted in [51]). Government checkers reviewing numerical codes assigned to categories of commodities by business importers introduced nearly as many errors as they detected, reducing the overall error rate only from 8% to 7%.

More generally, Naus [51] pointed out that there are three ways to proceed if one suspects that an observation may be erroneous: resolution, deletion, and imputation. (If only extreme values provoke suspicion, some would argue that a fourth possibility — merely reporting the frequency of such observations without altering them — is preferable on the grounds that any corrective action introduces bias.) The first of these may involve a great deal of effort and expense, but maximizes the final data quality. The second costs nothing, but minimizes the amount of usable data. It would be worthwhile to have more information about the circumstances in which the third, imputation of corrected observations, is the most reasonable alternative.

Some level of data auditing by a party independent of investigators and sponsors clearly is needed for several reasons. As a recent, highly publicized example reminds us [52], there is the possibility of outright fraud. Certainly it is only by checking that one can discover where further clarification or training is needed. The question is how to meet these objectives with a focused program involving inspection of a small fraction of the data.

In the clinical trials setting, error detection early enough that correction is still possible inevitably means redundant processing wherever the potential for mistakes is significant. It will be necessary to analyze and monitor the entire data management environment in enough detail that one can determine where improved planning and training can reduce error rates effectively to zero even without double processing, and where they cannot. (The concept here, that inspection cannot produce data of high quality, is derived from the quality management and quality improvement ideas of Deming ([53]; see especially chapters 2, 3, 4, and 15).

An example of the latter case may be the decision to modify dose of study medication in light of an adverse experience on the part of the volunteer. While the protocol will contain guidelines for such occurrences, these guidelines do not cover all possibilities. If the success of the trial depends on consistency of decision making, the way to achieve it is to provide for independent assessment by two clinicians with comparable qualifications, followed if necessary by discussion to resolve disagreements.

To begin, double processing might be employed extensively, being reduced as evidence accumulates that particular steps are error free. To make the best use of the information available, formal tracking and analysis of error rates are desirable.

West and Winkler [54] proposed a method for estimating the number of errors remaining in a data set that consists of a set of independent observations made on the same binary variable. Discrepant pairs of observations are identifiable instances of error and provide a basis for estimating the error rates for the observers (or recorders). If both observers make the same error, it will not be discovered. In a Bayesian formulation, West and Winkler develop a technique for predicting the number of such errors remaining in the data set.

This is a promising approach, but it needs generalization in several directions to be really useful in multicenter clinical trials. First, most observations are not binary; for this and other reasons, one must consider a more elaborate model for the error probabilities. It is also desirable to account for the possibility of error rates that differ according to type and source of observation. From the viewpoint of managing resources, it would be important to construct methods that are useful in promptly detecting changes in error rates.

Discussion

Identification and use of surrogate markers, assessment of health-related quality of life, and treatment noncompliance will continue to require development of new clinical trial methodology in the context of AIDS clinical trials. New statistical methods for monitoring and assuring the quality of clinical trials data could perhaps have an even greater impact on clinical research, making it substantially more efficient — and not only in HIV/ AIDS research, of course.

Although there has not been space to discuss it at length, methodology has been developed to deal with other issues and types of data in AIDS clinical trials. Given the long time course of HIV disease, the evaluation of patient response generally involves repeated measurements over time. While a well-established methodology exists for failure-time endpoints, there is less consensus about how to analyze response *profiles* such as might be obtained from CD4 counts. The naive approach of conducting repeated tests over time introduces the problem of multiple comparisons and (especially given the lack of independence of such tests) tends to yield results that are difficult to interpret. A simple alternative discussed by Matthews et al. [55] uses a summary measure of the multiple responses for each individual. Useful functions for HIV laboratory markers have included the slope, area under the curve (AUC), and the mean. The measure can be chosen to capture aspect(s) of the response pattern of primary interest. This topic was further studied by Frison and Pocock [56].

Many AIDS clinical trials involve multiple agents: patients in an advanced disease state may be at risk for any of a number of opportunistic pathogens that may require different prophylaxes; the use of multiple anti-HIV drugs has been emphasized as a possible way to overcome the problems of resistance and toxicity that have hampered current monotherapies. Often a full factorial design must be ruled out. For example, ethical considerations may preclude a placebo group, or placebos may be allowed for some but not all active drugs in the study. Such constraints suggest the use of a *restricted* factorial design. Some implications of such a design for analysis of treatment effects were discussed by Byar [2].

The circumstances of the current state of AIDS treatments has motivated a rethinking of the desired goals and emphases of clinical trials. The long-term treatment of HIV-infected patients often entails many therapy changes and adjustments involving an assortment of therapeutic and prophylactic agents. Treatment thus represents a complex and reactive strategy that is difficult to capture or handle in a conventional 'explanatory' clinical trial. Many questions remain regarding optimal treatment strategies using available drugs; such effects will generally not be expected to be of great magnitude but may still be clinically important. Such considerations have led to the development of the concept of the large, simple trial (see [57,58]). Large, simple trials focus on the effectiveness of a treatment strategy and are characterized by broad entrance criteria and minimal, 'low-tech' data collection; these features allow larger samples and thus greater power to detect modest effects.

The rapidly changing understanding of AIDS continues to generate new challenges for the development of suitable methodology. Statisticians involved in AIDS clinical trials have drawn on experiences in other disease areas, notably cancer. Some of the circumstances of AIDS appear to be unique, but may reveal analogous problems in other research areas. We hope, therefore, that the insights and developments from AIDS are instructive and inspiring to statisticians working in other areas, and urge that the methodological challenges of AIDS continue to be met by the statistical community.

References

1. Fauci AS (1993). *The National Institutes of Health Five-Year Plan for HIV-Related Research*. Office of AIDS Research: Washington, DC.
2. Byar DP, Schoenfeld DA, Green SB, et al. (1990). Design considerations for AIDS trials. *N Engl J Med* 323:1343–1348.
3. Green SB, Ellenberg SS, Finkelstein DM, et al. (1990). Issues in the design of drug trials for AIDS. *Controlled Clin Trials* 11:80–87.
4. Ellenberg SS, Cooper E, Eigo J, et al. (1992). Studying treatments for AIDS: new challenges for clinical trials. *Controlled Clin Trials* 13:272–292.
5. Ellenberg SS, Finkelstein DM, Schoenfeld DA (1992). Statistical issues arising in AIDS clinical trials (with comments and rejoinder). *J Am Stat Assoc* 87:562–583.
6. Dixon DO, Rida WN, Fast PE, Hoth DF (1993). HIV vaccine trials: Some design issues including sample size calculation. *J AIDS* 6:485–496.
7. Ellenberg SS, Dixon DO (1994). Statistical issues in designing clinical trials of AIDS treatments and vaccines. *J Stat Planning Inf* 42:123–135.
8. Amato DA, Lagakos SW (1990). Considerations in the selection of end points for AIDS clinical trials. *J AIDS* 3 (Suppl 2):S64–S68.
9. Lagakos SW, Hoth DF (1992). Surrogate markers in AIDS: Where are we? Where are we going? *Ann Intern Med* 116:599–601.
10. Moss A (1990). Laboratory markers as potential surrogates for clinical outcomes in AIDS trials. *J AIDS* 3 (Suppl 2):S69–S71.
11. Mildvan D (1993). Clinical validation of virologic and immunologic assays. Presentation at the 17th AIDS Clinical Trials Group Meeting.
12. Machado SG, Gail MH, Ellenberg SS (1990). On the use of laboratory markers as surrogates for clinical endpoints in the evaluation of treatment for HIV infection. *J AIDS* 3:1065–1073.
13. Prentice RL (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat Med* 8:431–440.
14. Choi S, Lagakos SW, Schooley RT, Volberding PA (1993). CD4+ lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV infection taking zidovudine. *Ann Intern Med* 118:674–680.
15. Lin DY, Fischl MA, Schoenfeld DA (1993). Evaluating the role of CD4-lymphocyte counts as surrogate endpoints in human immunodeficiency virus clinical trials. *Stat Med* 12:835–842.
16. DeGruttola V, Wulfsohn M, Fischl MA, Tsiatis A (1993). Modeling the relationship between survival and CD4-lymphocytes in patients with AIDS and AIDS-related complex. *J AIDS* 6:359–365.
17. Concorde Coordinating Committee (1994). Concorde: MRC/ANRS randomised, double-blind controlled trial of immediate and deferred zidovudine in symptom-free HIV infection. *Lancet* 343:871–881.

18. Kosorok MR, Fleming TR (1993). Using surrogate failure time data to increase cost effectiveness in clinical trials. *Biometrika* 80:823–833.
19. Fleming T, Prentice R, Pepe M, Glidden D (1994). Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Stat Med* 13:955–968.
20. Pepe MS (1992). Inference using surrogate outcome data and a validation sample. *Biometrika* 79:344–365.
21. Pepe MS, Reilly M, Fleming TR (in press). Auxiliary outcome data and the mean score method. *J Stat Planning Inference.*
22. Finkelstein DM, Schoenfeld DA (1994). Analyzing survival in the presence of an auxiliary variable. *Stat Med* 13:1747–1754.
23. Guyatt G, Feeny D, Patrick D (1991). Issues in quality-of-life measurement in clinical trials. *Controlled Clin Trials* 12 (Suppl):81S–90S.
24. Fitzpatrick R, Fletcher A, Gore S, Jones D, Spiegelhalter D, Cox D (1992). Quality of life measures in health care. I: Applications and issues in assessment. *Br Med J* 305:1074–1077.
25. Cox DR, Fitzpatrick R, Gore SM, Spiegelhalter DJ, Fletcher AE, Jones DR (1992). Quality-of-life assessment: can we keep it simple (with discussion)? *J R Stat Soc A* 155: 353–393.
26. Williams JBW, Rabkin JG (1991). The concurrent validity of items in the quality of life index in a cohort of HIV-positive and HIV-negative gay men. *Controlled Clin Trials* 12 (Suppl):129S–141S.
27. Spitzer WO, Dobson AJ, Hall J, Chesterman E, Levi J, Shepherd R, Battista RN, Catchlove BR (1981). Measuring the quality of life of cancer patients: a concise QL-index for use by physicians. *J Chron Dis* 34:585–597.
28. Wu AW, Rubin HR, Mathews WC, Ware JE, Brysk LT, Hardy WD, Bozzette SA, Spector SA, Richman DD (1991). A health status questionnaire using 30 items from the medical outcomes study. Preliminary validation in persons with early HIV infection. *Med Care* 29:786–798.
29. Glasziou PP, Simes RJ, Gelber RD (1990). Quality adjusted survival analysis. *Stat Med* 9:1259–1276.
30. Gelber RD, Gelman RS, Goldhirsch A (1989). A quality of life oriented end point for comparing therapies. *Biometrics* 45:781–796.
31. Fischl MA, Richman DD, Hansen N, et al. (1990). The safety and efficacy of zidovudine (AZT) in the treatment of subjects with mildly symptomatic human immunodeficiency virus type 1 (HIV) infection. *Ann Intern Med* 112:727–737.
32. Gelber RD, Lenderking WR, Cotton DJ, et al. (1992). Quality-of-life evaluation in a clinical trial of zidovudine therapy in patients with mildly symptomatic HIV infection. *Ann Intern Med* 116:961–966.
33. Volberding PA, Lagakos SW, Koch MA, et al. (1990). Zidovudine in asymptomatic human immunodeficiency virus infection. *N Engl J Med* 322:941–949.
34. Lenderking WR, Gelber RD, Cotton DJ, Cole BF, Goldhirsch A, Volberding PA, Testa MA (1994). Evaluation of the quality of life associated with zidovudine treatment in asymptomatic human immunodeficiency virus infection. *N Engl J Med* 330:738–743.
35. Sackett DL, Snow JC (1979). The magnitude of compliance and noncompliance. In *Compliance in Health Care*, RB Haynes, DW Taylor, DL Sackett (eds.). Baltimore, MD, The Johns Hopkins University Press, 11–22.
36. Schork MA, Remington RD (1967). The determination of sample size in treatment-control comparisons for chronic disease studies in which dropout or non-adherence is a problem. *J Chron Dis* 20:233–239.
37. Lachin JM, Foulkes MA (1986). Evaluation of sample size and power for analysis of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics* 42:507–519.
38. Peduzzi P, Wittes J, Detre K, Holford T (1993). Analysis as-randomized and the problem

of non-adherence: an example for the Veterans Affairs randomized trial of coronary artery bypass surgery. *Stat Med* 12:1185–1195.

39. Lee YJ, Ellenberg JH, Hirtz DG, Nelson KB (1991). Analysis of clinical trials by treatment actually received: is it really an option? *Stat Med* 10:1595–1506.
40. May GS, DeMets DL, Friedman LM, et al. (1981). The randomized clinical trial: bias in analysis. *Circulation* 64:669–673.
41. Coronary Drug Research Group (1980). Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project. *N Engl J Med* 302:1038–1041.
42. Efron B, Feldman D (1991). Compliance as an explanatory variable in clinical trials. *J Am Stat Assoc* 86:9–26.
43. Mark SD, Robins JM (1993). A method for the analysis of randomized trials with compliance information: an application to the multiple risk factor intervention trial. *Controlled Clin Trials* 14:79–97.
44. Albert JM, DeMets DL (1994). On a model-based approach to estimating efficacy in clinical trials. *Stat Med* 13:2323–2335.
45. Robins JM, Tsiatis AA (1991). Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Commun Stat Theory Methods* 20:2609–2631.
46. Lagakos SW, Lim LL-Y, Robins JM (1990). Adjusting for early treatment termination in comparative clinical trials. *Stat Med* 9:1417–1424.
47. Spilker B (1991). *Guide to Clinical Trials*. New York: Raven Press.
48. Lim LL-Y (1992). Estimating compliance to study medication from serum drug levels: application to an AIDS clinical trial of zidovudine. *Biometrics* 48:619–630.
49. Richardson D, Liou S-H, Kahn JO (1993). Uric acid and didanosine compliance in AIDS clinical trials: an analysis of AIDS Clinical Trials Group protocols 116A and 116B/117. *J AIDS* 6:1212–1223.
50. Pritzker L, Ogus J, Hansen MH (1965). Computer editing methods — some applications and results. Proceedings 35th session, Belgrade. *Bull Int Stat Inst* 41:442–465.
51. Naus JI (1977). *Data Quality Control and Editing*. New York: Marcel Dekker.
52. Altmann L (1994). Report of fraud in a major breast cancer study. *New York Times*.
53. Deming WE (1986). *Out of the Crisis*. Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study.
54. West M, Winkler R (1991). Data base error trapping and prediction. A focus on Bayesian methods. *J Am Stat Assoc* 86:987–996.
55. Matthews JNS, Altman DG, Campbell MJ, Royston P (1990). Analysis of serial measurements in medical research. *Br Med J* 300:230–235.
56. Frison L, Pocock S (1992). Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Stat Med* 11:1685–1704.
57. Foulkes MA, Ellenberg SS (in preparation). Large simple trials of HIV therapies. In *AIDS Clinical Trials*, DM Finkelstein, DA Schoenfeld (eds).
58. Yusuf S, Collins R, Peto R (1984). Why do we need some large, simple trials? *Stat Med* 3:409–420.

3. Recent developments in the design of phase II clinical trials

Peter F. Thall and Richard M. Simon

Introduction

Clinical trials of new medical treatments may be classified into three successive phases. *Phase I* trials typically are small pilot studies to determine the therapeutic dose of a drug, biological agent, radiation schedule, or a combination of these regimens (cf. [1]). In cancer therapeutics, the underlying idea is that a higher dose of the therapeutic agent kills more cancer cells but also is more likely to harm and possibly kill the patient. Consequently, toxicity is the usual criterion for determining a maximum tolerable dose (MTD), and most phase I cancer trials involve very small groups of patients, usually three to six patients per dose, with each successive group receiving a higher dose until it is likely that the MTD has been reached. A more refined approach that continually updates an estimate of the probability of toxicity has also been proposed by O'Quigley, Pepe and Fisher [2].

Once a dose and schedule of a new experimental regimen E have been determined, its therapeutic efficacy is evaluated in a *phase II* trial. Phase II trials are usually single-arm studies involving roughly $n = 14$ to 90 patients treated with E, with n usually well under 60. These studies typically are carried out within a single institution and are most prominent in clinical environments where there are many new treatments to be tested. The primary goal is to determine whether E has a level of antidisease activity sufficiently promising to warrant its evaluation in a subsequent phase III trial (described below). Phase II results also frequently serve as the basis for additional single-arm studies involving E in other combination regimens or dosage schedules. The main statistical objective of a phase II trial thus is to provide an estimator of the response rate associated with E (cf. [3]). Treatment success generally is characterized by a binary patient response, such as 50% or more shrinkage of a solid tumor or complete remission of leukemia, and the scientific focus is p, the probability of response with E. Patient response usually is defined over a relatively short time period in phase II, based on the underlying idea that short-term response is a necessary precursor to improved long-term survival and reduction in morbidity. Phase II trials are important because they are the primary means of selecting

Peter F. Thall (ed.), RECENT ADVANCES IN CLINICAL TRIAL DESIGN AND ANALYSIS.
Copyright © 1995 Kluwer Academic Publishers, Boston. All rights reserved. ISBN 978-1-4613-5830-5.

treatments for phase III evaluation, and moreover, many patients receive treatment within the context of a phase II trial.

The ultimate standard for evaluation of medical treatments is the randomized comparative *phase III* clinical trial. Phase III trials generally are large, multi-institutional studies with treatments evaluated in terms of long-term patient response, such as survival or time to disease progression. Phase III trials are designed and conducted to evaluate the effectiveness of a treatment relative to an appropriate control and with regard to endpoints that represent patient benefit, such as survival. To achieve such objectives, the trial design is based on statistical tests of one or more hypotheses and may require approximate balance and minimal sample size within important patient subgroups. Because they are larger and of longer duration than phase II trials, and typically involve multiple institutions, phase III trials are usually much more costly and logistically complicated. The results of phase III trials are broadly disseminated within the medical community and form the basis for changes and advances in general medical practice.

The simplest phase II design is a single-arm, single-stage trial in which n patients are treated with E. The data consist of the random variable Y_n, namely, the number of successes after n patients are evaluated, which is binomial in n and p. The sample size is determined so that, given a fixed standard rate p_0 that is of no clinical interest, a test of H_0: $p \leqslant p_0$ versus H_1: $p \geqslant p_1$ has type I error probability (significance level) $\leqslant \alpha$ and type II error probability $\leqslant \beta$ for a given target response probability $p_1 = p_0 + \delta$. The test is determined by a cutoff r, with H_0 rejected if $Y_n \geqslant r$ and H_1 rejected if $Y_n < r$. A type I error occurs if it is concluded that E is promising compared to standard therapy, i.e., if H_1 is accepted, when in fact $p \leqslant p_0$. The consequences of this are that an uninteresting or even inferior treatment is likely to become the basis for a phase III trial, and that if future phase II trials using a combination therapy based on E are conducted, the patients in those trials will be treated with an inferior agent while phase II trials of other potentially promising new treatments are delayed. A type II error occurs if it is concluded that E is not promising compared to standard therapy, i.e., if H_0 is accepted when in fact $p \geqslant p_0 + \delta$. The power of the test is $1 - \beta$, the probability of correctly accepting H_1 when E really has success rate $p_0 + \delta$. The consequence of a type II error is that a promising treatment has been lost or its detection delayed. The required sample size n and test cutoff r are determined by specifying α, β, p_0, and δ. Since there is a trade-off between type I and type II error, in practice typically $(\alpha,\beta) = (0.10,0.10)$, $(0.05,0.20)$, or $(0.05,0.10)$. We shall refer to α and β, and more generally any parameters that describe a design's behavior, as its *operating characteristics*.

Smaller treatment advances δ are harder to detect, i.e., they require a larger sample size for given p_0, α, and β. A very large δ requires a trivially small sample size, i.e., it is easy to detect a large treatment advance. Reasonable values are thus $\delta = 0.15$ to 0.20, since $\delta < 0.15$ usually leads to unrealistically large n while $\delta > 0.20$ leads to a trial yielding very little information about E and in many cases is intellectually dishonest. Parameters of some typical single-stage designs are given in table 1.

An alternative to designing a single-stage trial in terms of hypothesis testing, which is a formal method for deciding whether E is promising compared to the fixed-standard success probability p_0, is to choose n to obtain a confidence interval of given width and level (coverage probability) to estimate p. A good approximate confidence interval, due to Ghosh [4], is

$$\frac{\hat{p} + A/2 \pm z\{\hat{p}(1 - \hat{p})/n + A/(4n)\}^{1/2}}{1 + A},$$

where $\hat{p} = Y_n/n$, $z = 1.645$, 1.96, or 2.576 for a 90%, 95%, or 99% coverage probability, respectively, and $A = z^2/n$. The exact binomial confidence interval of Clopper and Pearson [5] also may be used, although the above approximation is quite adequate for planning purposes. An important caveat is that the commonly used approximate interval $\hat{p} \pm z\{\hat{p}(1 - \hat{p})/n\}^{1/2}$ is rather inaccurate for many values of n and p encountered in phase II trials [4] and is not recommended. Table 2 gives the sample sizes needed to obtain 90% or 95% confidence intervals for p of given width, based on values of $\hat{p}$ from 0.20 to 0.50. The sample sizes for $\hat{p} = 0.50 + \Delta$ and $0.50 - \Delta$ are identical. For example, if it is anticipated that the empirical rate Y_n/n will be approximately 0.30 or 0.70, then a sample of 34 patients is required to obtain a 90% confidence interval for p having width at most 0.25. Given an observed rate of 10/34, one could be 90% certain that the true success probability of E is somewhere between 0.185 and 0.434.

Although the single-stage design is easy to understand and implement, it has several severe practical limitations. Each of the designs described in the following sections was created to address one or more of the following problems.

Table 1. Single-stage designs Conclude $p \geqslant p_1$ at level α and power $1 - \beta$ if $Y_n/n \geqslant r/n$.

			(α,β)		
δ	p_0	p_1	(0.10,0.10)	(0.05,0.20)	(0.05,0.10)
0.20	0.10	0.30	5/25	6/25	7/33
	0.20	0.40	11/36	12/35	15/47
	0.30	0.50	16/39	17/39	22/53
	0.40	0.60	21/41	23/42	29/56
0.15	0.10	0.25	7/40	8/40	10/55
	0.20	0.35	17/61	17/56	22/77
	0.30	0.45	27/71	27/67	36/93
	0.40	0.55	36/75	36/71	46/94

Table 2. Single-stage n to obtain confidence interval of given level and width $\leq W$

		Anticipated $\hat{p} = Y_n/n$			
Level	W	0.20	0.30	0.40	0.50
90%	0.20	44	55	63	66
	0.25	26	34	40	42
	0.30	19	24	26	28
95%	0.20	64	78	89	94
	0.25	39	48	56	58
	0.30	26	33	38	40

1. The most serious limitation of the single-stage design is that it ignores all data prior to observation of Y_n, and in particular has no provision for early termination if the interim observed response rate is unacceptably low. For example, if $p_0 = 0.30$ is the established response rate with standard treatment and E also has rate $p = 0.30$, then an initial run of 12 failures should occur with probability 0.014, and if $p > 0.30$ then such a run has probability close to 0. Most clinicians would be strongly inclined to discontinue use of E at or before this point, especially in trials of treatments for rapidly fatal diseases or other circumstances where early failure increases morbidity or reduces survival. Designs with early stopping rules address this problem (cf. [6–14]).

2. Reporting results of a phase II trial entails augmenting or replacing significance test results with a confidence interval for p, since the real goal of a phase II trial is estimation [3]. If rules for early stopping are included in the design, however, then computation of the confidence interval for p based on the final data must account for the fact that the trial continued through its intermediate stages, since the usual unadjusted confidence intervals are biased in this case. Methods for computing a confidence interval for p after a multistage trial have been given by numerous authors, including Jennison and Turnbull [15], Tsiatis, Rosner, and Mehta [16], Atkinson and Brown [17], and Duffy and Santner [18].

3. Another problem, addressed by Thall and Simon [19], is that p_0 often is estimated from historical data and hence is a statistic $\hat{p}_0$, not a fixed value. Since this estimator has an associated variance, the usual test statistic $Y_n/n - \hat{p}_0$ has variance $p(1 - p)/n + \text{var}(\hat{p}_0)$. The sample size computation that ignores $\text{var}(\hat{p}_0)$ is incorrect, and the actual type I and type II error rates are larger than their nominal values.

4. In some settings several new treatments may be ready simultaneously for phase II evaluation. The question then arises of whether to carry out a sequence of single-arm trials or one randomized trial, and in either case strategies are needed for prioritizing treatments and for selecting one or

more promising treatments from those tested. Several approaches to this general problem have been proposed. Simon, Wittes, and Ellenberg [20] propose a randomized phase II trial; Whitehead [21] proposes a combined phase II–III strategy; Thall, Simon, and Ellenberg [22,23] propose 'select then test' designs for comparing the best of several experimental treatments to a standard; and Strauss and Simon [24] examine properties of a sequence of 'play the winner' randomized phase II trials.

5. The assumption that patient response can be characterized effectively by a single variable is rather strong, even for short-term response, and it may be necessary to monitor more than one patient outcome. For example, in most cancer chemotherapy trials, toxicity is an important issue, and it is highly desirable to have an early stopping rule to protect future patients from unacceptably high rates of toxicity. Many phase II trials include such a rule either formally or informally in their protocols, but they ignore the interdependence between toxicity and response in the design. Designs accounting for multiple outcomes have been proposed by Etzioni and Pepe [25] and Thall, Simon, and Estey [26].

6. Patient-to-patient variability is often high, even in clinical trials with very specific entry criteria. Since phase II trials are relatively small, a study with an unusually high proportion of either poor-prognosis or good-prognosis patients may give a misleadingly pessimistic or optimistic indication of how E would behave in the general patient population.

7. Although most phase II designs regard treatment response rate p as a fixed unknown quantity, many clinicians regard p as random. For example, when asked to specify p_0, the clinician may respond by giving a range rather than a single value, and may even describe the probability distribution of p_0 within that range. In such circumstances, a Bayesian design, based on random values of p_0 and p, may be more appropriate. Bayesian phase II designs have been proposed by Sylvester and Staquet [27,28] Sylvester [29], Etzioni and Pepe [25], and Thall and Simon [12–14], and Thall, Simon and Estey [30].

Refinements of the phase I–II–III paradigm

When the best available therapy has little or no effect against the disease, the phase II trial's objective is to determine whether E has any antidisease activity at all. This is a phase IIA trial. Since $p_0 = 0$ or possibly 0.05 in this case, type II error is the main consideration. Gehan [6] proposed the first phase IIA design, a two-stage design in which n_1 patients are treated at stage 1, the trial is stopped if $Y_{n_1} = 0$, and an additional n_2 patients are treated in stage 2 if $Y_{n_1} > 0$. The stage 1 sample size is chosen to control type II error, specifically $n_1 \geqslant \log(\beta)/\log(1 - p_1)$ for targeted success rate p_1. The stage 2 sample size is chosen to obtain $\hat{p}$ having standard error no larger than a given magnitude, and n_2 also depends on Y_{n_1}. For example, if $\beta = 0.05$ and

$p_1 = 0.20$, then $n_1 = 14$ patients are required at stage 1. If $Y_{14} > 0$, then to obtain an estimate of p having standard error 0.10 requires $n_2 = 1, 6, 9$, or 11 if Y_{14} is 1, 2, 3, or $\geqslant 4$, respectively.

When there exists a standard treatment, say S, having some level of activity (i.e., when $p_0 > 0$), then the goal is to identify new treatments that are promising compared to p_0. This is a phase IIB trial. In this case, there are compelling data, arising from clinical trials or in vitro testing, indicating that E is likely to be active at a level exceeding p_0. An important consideration in IIB trials is that it is clinically undesirable to continue a trial of an experimental treatment that proves to be not promising compared to S. For example, when $p_0 = 0.40$ and $p_1 = 0.55$, if interim trial results strongly indicate that $p < 0.40$, then it is unethical to continue; if it is likely that $0.40 \leqslant p < 0.55$, then it may be desirable to terminate the trial to make way for other, potentially more promising new treatments. It is also important to recognize the comparative aspect of phase IIB trials, which may lead to formal use of historical data on S in the evaluation of E, and possibly to a randomized trial [19]. This issue will be discussed below.

If several new treatments are simultaneously available for phase II testing, then the problem of choosing among them arises. Since the number of patients in any clinic is limited, this situation frequently occurs in institutions with high levels of research activity in growth factors or pharmacologic agents. Thall and Estey [30] propose a pre-phase II Bayesian strategy in which patients having a prognosis more favorable than that of phase I patients but less favorable than that of the target group of the subsequent phase II trial are randomized among several experimental treatments. The response rate distribution in each treatment arm is updated continually during the trial and is compared to early termination cutoffs, and the best final treatment must satisfy a minimal posterior efficacy criterion before it is evaluated in a subsequent phase II trial. This type of study, the phase I.5 trial, bridges the gap between phase I and phase IIB. It provides an ethical means of giving poor-prognosis patients experimental treatments while replacing the usual informal pre-phase II treatment selection process with a fair comparison formally based on a combination of prior opinion and clinical data.

As an example, a phase I.5 trial might be carried out in patients who have acute myelogenous leukemia (AML) with $\geqslant 1$ prior relapse and poor-prognosis cytogenetic characteristics, in order to select a treatment for phase II testing in untreated AML patients who have good-prognosis cytogenetics. If the accrual rate is 40 per year in the poor-prognosis group, then a phase I.5 trial of three treatments with up to 10 patients per treatment arm could be carried out in nine months. Assuming a prior mean response rate of 0.40 for all three arms, Thall and Estey [30] recommend a design in which a treatment arm is terminated if there are 0 responses in the first 4 patients; otherwise, 10 patients are accrued in that arm. The best treatment, among those not terminated, must have $\geqslant 4$ responses to be selected for the phase II trial.

The response rates obtained in different phase II trials of the same treatment often vary widely. Simon, Wittes, and Ellenberg [20] cite a number of factors as the sources of this variability, including patient selection, definition of response, interobserver variability in response evaluation, drug dosage and schedule, reporting procedures, and sample size. To deal with these problems, these authors propose randomizing patients among several experimental treatments in phase II, with ranking and selection methods rather than hypothesis testing used to evaluate treatments. They recommend the use of conventional phase II sample sizes and early stopping criteria in each treatment arm, and that a standard treatment arm not be included. Specifically, they propose that sample size be computed to ensure that, if one group of treatments has response rate $p_0 + \delta$ and the rest have rate p_0, then a 'select the best' strategy will choose one of the superior treatments with a desired probability. For example, if $p_0 = 0.20$ and $\delta = 0.15$, then 44 patients in each of three arms will ensure a 90% chance of choosing a treatment with response rate 0.35.

Strategies for phase II evaluation of new treatments that become available sequentially over time have been considered by Whitehead [31] and by Strauss and Simon [24]. Whitehead is motivated in part by the desire to examine the properties of small sample sizes for phase II studies. He assumes that the success rates of the experimental treatments are random and may be considered as independent draws from a beta prior distribution. Given N equal to the total number of patients for all the trials, he derives the number of trials k and number of patients per trial n that maximize the expected success probability $E(\pi)$ of the selected treatment, subject to $nk = N$. For example, if $N = 60$ and the mean experimental success rate is 0.20, then depending upon prior variability, the optimal integer values of (n, k) and $E(\pi)$ vary from (4,15) with $E(\pi) = 0.426$, to (6,10) with $E(\pi) = 0.292$.

Strauss and Simon [24] study properties of a sequence of comparative phase II trials. At each of k stages, $2n$ patients are randomized between a new experimental treatment and the better of the two treatments from the previous stage, starting with a known standard S at stage 1. The better of the two treatments at each stage (the 'winner') thus becomes the new standard, and is then compared to the next experimental treatment. The goal is to select a single treatment for phase III evaluation. Similar to Whitehead [31], Strauss and Simon assume that the success probabilities of the experimental treatments are independent draws from a beta prior distribution, either with fixed mean equal to that of S or with distribution adapted to the data in that its mean equals that of the latest winner. This approach, however, is more robust against time trends in the selection of patients. Given a total of $N = nk$ patients, they examine the manner in which the expected success probability $E(p)$ of the final selected treatment varies with n, k, and N. They identify conditions under which such a sequence of phase II trials is more likely than a single phase II trial to identify a promising experimental treatment.

Whitehead [21] also proposes an integrated approach to the problem of evaluating several new treatments. A sequence of single-arm phase II trials is conducted; the most promising experimental treatment among them is selected, and it is then compared to the standard in a phase III trial. Assuming that the success rates of the experimental treatments are random and may be considered as independent draws from a beta prior distribution, Whitehead derives strategies for dividing patients between the two phases (given the number of phase II trials and the total number of patients) that maximize the probability π of obtaining a significant result in the phase III trial. For example, if $N = 300$ patients are available and there are five new agents to be tested, then allocating 18 patients to each of the five phase II trials and 210 to phase III ensures that $\pi = 0.52$. If instead $N = 500$, then the optimal allocation is 31 with 345 in phase III, which ensures that $\pi = 0.63$. Whitehead notes that, when using this strategy, the main trade-off is between the total numbers of patients allocated to the two phases.

Some practical considerations

Because phase II trials are developmental, their design and conduct must include several ethical and logistical considerations. These include the appropriateness of treating patients with E, the relevance of the trial within the larger context of treatment development, the patient accrual rate, definition of patient response, and the monetary cost of the trial. In any phase II setting, a priori there must be a reasonable basis for the belief that E may provide an improvement over the standard, whether $p_0 = 0$ or $p_0 > 0$. If in the course of the trial it becomes clear that this is unlikely, then it may be desirable to terminate early, and here the unavoidable conflict between type I and type II error comes into play. The trade-off is between protecting patients from an ineffective or dangerous experimental regimen and risking the loss of a treatment advance. If an adverse outcome, such as toxicity, is monitored along with the usual efficacy outcome, then an alternative goal may be to decrease the adverse event rate while maintaining a given response rate. Designs which monitor multiple events, such as response and toxicity, are discussed in a later section.

Ethical considerations are most pressing for rapidly fatal diseases, and the standards of clinical conduct for such diseases may provide a basis for analogous decisions in less extreme circumstances. The desirability of a particular treatment E in a phase II trial must be assessed from the viewpoints of the individual patient, all patients in the trial taken as a group, and future patients after the trial is completed. A general consideration is that patients are more likely to choose a physician rather than a treatment and to rely on their physician's advice regarding treatment choice. The centuries-old process of entrusting one's life and well-being to one's physician is a fundamental part of medicine, informed consent notwithstanding. Thus, the trial must be designed so that trial objectives and individual patient benefit are not in conflict. The situation is most desperate in phase IIA trials of treatments for rapidly fatal diseases for which no effective treatment exists. The trade-off for both the individual patient and for the trial is between the risk of adverse treatment effects and the likelihood of any therapeutic benefit. For nonfatal diseases, the potential severity of adverse effects first must be weighed against the effects of the disease itself, and it is inappropriate to conduct a trial of E if its effects are likely to be worse than those of the disease. Phase IIB trials often evaluate combination therapies whose components are already known to have antidisease activity. Consequently, a new combination regimen with an activity level below that of the standard is usually not promising for future development. Two exceptions are a trial in which a reduced likelihood of early response may be an acceptable trade-off for improved overall survival, and a trial in which the real goal is to reduce toxicity and a small reduction in response rate is considered an acceptable trade-off. Examples of such trials are given in a later section.

Patient accrual and monetary cost are absolute limits on the size of any clinical trial. If either the number of patients or the available resources are insufficient to achieve initial goals, then a smaller trial may be appropriate. However, the magnitudes of α and β and the reliability of the final estimate of p should be kept in mind when reducing sample size due to low accrual rate or limited resources. The results of very small trials often are of limited value and, due to their high variability, are potentially misleading. If resources are inadequate to conduct a trial that will produce useful results, then it is inappropriate to conduct the trial.

A simple but critical issue in trial design and conduct is definition of patient outcome. For example, in AML, treatment response is typically complete remission (CR), which is defined in terms of several parameters (e.g., blast count, platelet recovery, white cell count, etc.), as measured within a given timeframe. It is essential that CR be defined formally in the protocol and that, however CR is defined, all clinicians involved in the trial adhere to that definition. Otherwise, one clinician's CR may be another's failure, which renders the recorded trial results virtually meaningless. The same considerations apply to definition of adverse outcomes, since there are various grades of toxicity, etc. This problem is potentially more severe in multi-institutional phase II trials; hence, an even stronger effort must be made to define and score patient outcomes consistently.

Short-term response in a phase II trial is used as the measure of treatment effect. For solid tumors, however, partial response often is not a validated measure of patient benefit. In general, the comparison of survival between responders and nonresponders is not valid for demonstrating that treatment has extended survival for responders [32]. Because response is often viewed as a necessary but not sufficient condition for extending survival, response may be used in phase II trials for screening promising treatments. To evaluate the effectiveness of a regimen in prolonging survival, however, a phase III trial of survival is required.

Historical data and Bayesian designs

Most phase II trials evaluate one or more new treatments relative to a standard therapy S; hence, they are inherently comparative, even though a standard treatment arm usually is not included. In designing the single-stage, single-arm trial described in the introduction to this chapter, a common practice is to assume that p_0 is a known constant (and hence that the statistic $\hat{p} - p_0 = (Y_n/n) - p_0$ has variance $\text{var}(\hat{p}) = p(1 - p)/n$) and to determine n to obtain a test of $p = p_0$ versus $p = p_0 + \delta$ having given type I and type II error rates α and β. For phase IIB trials, where p_0 represents the activity level of available regimens, the numerical value of p_0 used in this computation is often a statistical estimate $\hat{p}_0$ based on historical data, rather than a known constant. The empirical difference $\hat{p}_1 - \hat{p}_0$, which is the basis for the test, is thus the difference between two statistics and has variance larger than the assumed $p(1 - p)/n$. Consequently, the sample size computed under a model ignoring the fact that $\hat{p}_0$ is a statistic is incorrect. This common practice may be due to the belief that the variability of $\hat{p}_0$ is of no practical consequence or to the absence of a theoretical basis and associated statistical software for computing sample sizes correctly.

Thall and Simon [19] derive optimal single-stage phase II designs that incorporate historical data from one or more trials of S and account for the variability inherent in $\hat{p}_0$. They consider both binary and normally distributed responses. Because the variability between historical pilot studies sometimes exceeds what is predicted by a binomial model for binary responses, they use a beta-binomial model to account for possible extrabinomial variation. Their results indicate that it is sometimes best to randomize a proportion of patients to S, and they derive the total sample size and optimal proportions for allocation to E and S that minimize $\text{var}(\hat{p}_1 - \hat{p}_0)$. Their results indicate that an unbalanced randomization may be superior to a single-arm trial of E alone, and that ignoring $\text{var}(\hat{p}_0)$ may lead to trials with actual values of α and β much higher than their nominal values. For example, consider a trial in which $\hat{p}_0 = 0.20$ is based on three historical trials of 20 patients each. To obtain a test that detects an improvement of $\delta = 0.20$, i.e., for alternative $p_1 = 0.40$, with $\alpha = 0.05$ and $\beta = 0.20$, the optimal design requires 85 patients with 27 allocated to S and 58 to E. If the variability in $\hat{p}_0$ is ignored and a single-arm trial of E is conducted, the standard computation yields $n = 35$, and the resulting test will have actual $\alpha = 0.14$ and $\beta = 0.27$. Since the numerical computations to incorporate the historical data and obtain the optimal design are somewhat complicated, a menudriven computer program written in Splus has been made available.

The above method for dealing with the variability of an estimate of p_0 may be regarded as a particular approach to a more general problem. Given that in a phase II trial the success rate of E ultimately must be compared to that of S, and that uncertainty regarding the response rate of S will always exist, the general problem is to account for this uncertainty when planning the trial and interpreting its results. A different statistical approach is based on the *Bayesian* framework, in which the success probabilities of E and S are regarded as random rather than fixed parameters. To underscore this distinction, we denote the random response probabilities by θ_S and θ_E. Although the theoretical basis for Bayesian methods is well established, practical methods for clinical trials have been proposed only recently, notably by Freedman and Spiegelhalter [33,34], Spiegelhalter and Freedman [35,36], Racine et al. [37], and Berry [38,39].

Sylvester and Staquet [28] and Sylvester [29] propose decision-theoretic Bayesian methods for phase II clinical trials. They optimize the sample size and decision cutoff of a single-stage design where n is fixed, to determine whether a new drug is active, by minimizing the Bayes risk. Their approach assumes that $\mathrm{P_r}\,[\theta_E = p_1] = 1 - \mathrm{P_r}\,[\theta_E = p_2]$, with $p_2 > p_1$, where p_2 and p_1 are response rates at which E would and would not be considered promising, respectively — i.e., they assume that θ_E may take on two possible values.

Herson [7] proposes the use of *predictive probability* (PP) as a criterion for early termination of phase II trials to minimize the number of patients exposed to an ineffective therapy. The PP of an event, such as concluding that E is or is not promising according to some decision rule, is the conditional probability of that event given the current data, computed by first averaging over the prior distributions of the parameters, which are θ_S and θ_E in the present context. Mehta and Cain [9] provide charts of early stopping rules based on the posterior probability of $[\theta_E > p_0]$, where p_0 is a fixed level at which E would be considered active.

Palmer [40] proposes a Bayesian procedure for identifying the best of three treatments E_1,E_2,E_3. He assumes that their respective success probabilities are $\pi_1 = (a,b,b)$, $\pi_2 = (b,a,b)$, or $\pi_3 = (b,b,a)$ with prior probability 1/3 each, where $b < a$ are known fixed standards, analogous to p_0 and $p_0 + \delta$ in the hypothesis-testing context. Given a maximum sample size N, patients are first randomized among the treatments in triplets, and based on the posterior probabilities of $\{\pi_1,\pi_2,\pi_3\}$ the worst treatment may be dropped. Patients are then randomized between the two remaining treatments in pairs, and the worse of the two is subsequently dropped based on the posterior distribution. The optimality criterion is to maximize the expected number of future treatment successes. Palmer gives an idealized example in which the respective true response rates of E_1,E_2,E_3 are 0.40, 0.35, 0.31, and $N = 300$. At the first stage, 42 patients are randomized in triplets before E_3 is rejected, and then 58 more patients are randomized in pairs between E_1 and E_2 before E_1 is chosen. As with any selection procedure, this method is subject to the error of choosing an inferior treatment, and Palmer provides numerical tables of operating characteristics.

Thall and Simon [12–14] present a Bayesian approach to phase II clinical trials in which patient response is binary and the accumulating data are monitored continuously. Their designs require an informative beta prior for θ_S, a flat or weakly informative beta prior for θ_E, a targeted improvement for θ_E over θ_S, and lower and upper bounds m and M on the allowable sample size. The maximum sample size M is chosen to obtain a given level of reliability in the posterior distribution of θ_E. Depending upon the specific objectives, the posterior distribution of θ_E is updated when each patient response is observed. The trial may be terminated if E is shown with high posterior probability to be either promising or not promising compared to S, or if the predictive probability of either conclusion is small. Otherwise, the trial continues. Although the framework for determining early termination bounds and M is Bayesian, the operating characteristics of the design are evaluated using frequentist criteria, and the design parameters are determined on that basis. Since the trial may be terminated early on the basis of interim results, the sample size is random and on average is smaller than M. This is the case for all designs with interim stopping rules, including multistage designs.

For example, suppose that the prior on θ_S has mean $\mu_S = 0.30$ and $W_{S,90} = 0.20$, i.e., the width of the 90% central prior interval for θ_S is 0.20 (formally, $P_r[0.20 < \theta_S < 0.40] = 0.90$). This corresponds to a beta distribution with parameters 16.62 and 38.78, which might arise from a previous study of S with roughly 55 patients and a 30% response rate. To obtain a posterior distribution such that $P_r[0.40 < \theta_E < 0.60 \mid Y_M] = 0.90$ requires $M = 65$ patients. This would ensure that once the trial is completed, one may be 90% certain that the success rate with E is within 0.10 of its mean value. Monitoring begins at $m = 10$ patients. For a targeted improvement of 0.20, the decision criteria after the nth patient outcome is observed are to stop the trial and declare E promising compared to S if $P_r[\theta_S < \theta_E \mid Y_n] > 0.95$, or to stop the trial and declare E not promising compared to S if $P_r[\theta_S + 0.20 < \theta_E \mid Y_n] < 0.05$, and otherwise to continue to accrue patients up to the maximum of 65. These criteria yield upper and lower stopping boundaries U_n and L_n for $n = m, \ldots, M$ such that E is declared promising if $Y_n \geqslant U_n$, E is declared not promising if $Y_n \leqslant L_n$, and the trial continues if $L_n < Y_n < U_n$. In the example, at the tenth outcome $(L_{10}, U_{10}) = (2,6)$, so the trial is stopped and E is declared not promising if $Y_{10} \leqslant 2$, E is declared promising if $Y_{10} \geqslant 6$, and the trail continues if $2 < Y_{10} < 6$. Likewise, $(L_{11}, U_{11}) = (2,7)$, $(L_{12}, U_{12}) = (3,7)$, etc.; so in practice, once the stopping rules are established, conduct of the trial is straightforward.

The operating characteristics of the design may be evaluated by computing the probabilities of declaring E promising, declaring E not promising, or accruing all 65 patients without either conclusion under fixed values p_E of the response rate of E. Important values are $p_E = \mu_S$ and $p_E = \mu_S + \delta$, the mean standard and targeted success rates, respectively. In the example, if the true success probability of E is the standard mean rate 0.30, then under this design the trial is terminated early and E is declared not promising with probability $p_- = 0.88$; if the true success probability of E is 0.50, then E is declared promising with probability $p_+ = 0.84$. In either case, the median sample size is 12.

An alternative design stops early only if E is not promising compared to S, and does not stop early if E is promising. This design would be preferred when it is desirable to continue the trial if the new treatment is promising rather than to terminate it early. With this design, if E has true success rate 0.30, then $p_- = 0.94$; if E has true success rate 0.50, then the design accrues all 65 patients with probability 0.85. In practice, p_+ and p_- are computed when planning the trial, and the values of p_L, p_U, δ, m, or M are modified as appropriate in order to obtain a design with desirable operating characteristics. Since the numerical computations necessary to implement this design are quite complicated, a menu-driven computer program written in Splus has been made available.

Multistage designs

Designs that provide criteria for early termination based on each outcome Y_n may be regarded as extreme versions of multistage designs, which provide early stopping rules at one or more interim points in the trial. Schultz et al. [41] and Fleming [8] provide a general multiple-testing framework for phase II trials in which n_j patients are accrued at the jth stage, $j = 1, \ldots, K$, and a decision is made to stop the trial or continue based on a test of H_0: $p \leqslant p_0$ versus H_1: $p \geqslant p_1$. Let $(a_1, \ldots, a_K)$ and $(r_1, \ldots, r_K)$ be sequences of lower and upper test cutoffs. At stage j, H_1 is rejected and the trial is terminated if $S_j = Y_{n_1+\ldots+n_j} \leqslant a_j$, H_0 is rejected and the trial is terminated if $S_j \geqslant r_j$, and the trial continues to the next stage if $a_j < S_j < r_j$. If the trial continues to the Kth (final) stage, then one of the two hypotheses must be rejected; hence, $a_K = r_K - 1$. The maximum sample size $M = n_1 + \ldots + n_K$ and test cutoffs must be chosen to provide overall test error rates α and β. The actual sample size N is thus random, taking on possible values n_1,$n_1 + n_2, \ldots, M$, depending upon the interim test results.

Fleming [8] provides an explicit method for determining the test cutoffs, although the number of stages and division of patients among the stages are somewhat arbitrary, aside from the error rate constraints. For example, a Fleming design to test $p \leqslant 0.30$ versus $p \geqslant 0.50$ with $\alpha = 0.05$ and $\beta = 0.11$ may be conducted with three stages of sizes $n_1 = 20$ and $n_2 = n_3 = 15$, and test cutoffs $(L_{20}, U_{20}) = (5,12)$, $(L_{35}, U_{35}) = (12,17)$ and $(L_{50}, U_{50}) = (20,21)$. If $p = 0.30$, the expected number of patients under this design is $E_0(N) =$

35.5. A two-stage Fleming design with the same size and power has $n_1 = n_2 = 25$ with test cutoffs (7,14) and (20,21), and null expected sample size $E_0(N) = 40.8$.

Therneau, Wieand, and Chang [42] provide an enumeration algorithm that derives Fleming designs for given $K, n_1, \ldots, n_K, \alpha, \beta, p_0$, and p_1 that are optimal in that they have minimal expected sample size. In practice, this algorithm may easily be extended to determine optimal interim sample sizes $n_1, \ldots, n_K$, as well, and a computer program to derive the optimal designs is available. Since much of the advantage of multistage designs over a single-stage trial is achieved for $K = 2$, Simon [11] derives two-stage designs that either (1) minimize $E_0(N)$ (the *optimal* design) or (2) minimize the maximum sample size M (the *minimax* design) for given α, β, p_0, and p_1. An important distinction between the two-stage version of the Fleming design and Simon's designs is that the latter allow only rejection of H_1 or continuation but not rejection of H_0 at the interim test. For the hypotheses considered above, the optimal Simon two-stage design requires $n_1 = 24$ patients initially, stops after stage 1 and rejects H_1 if $Y_{24} \leqslant 8$, and otherwise accrues an additional $n_2 = 39$ patients with final test cutoffs (24,25). The corresponding minimax design has $n_1 = 24$ with stage 1 cutoff 7 and $n_2 = 29$ with final test cutoffs (21,22). For these designs, $E_0(N) = 34.7$ and 36.6, respectively. Simon (1989) tabulates design parameters and operating characteristics for a broad range of parameter values, and a computer program to obtain these values is also available.

Garnsey-Ensign et al. [43] provide an optimal three-stage design that is essentially a combination of the Gehan [6] and optimal Simon [11] designs. At stage 1, the design stops with rejection of H_1 if there is an initial run of n_1 failures; otherwise, it continues to stage 2 and (possibly) stage 3, which have decision rules analogous to those in stages 1 and 2 of Simon's designs. Rejection of H_1 is thus possible at any stage, but H_1 may be accepted only at the final test. The design is optimal in that $E_0(N)$ is minimized for given α, β, p_0, and p_1, subject to the constraint $n_1 \geqslant 5$. To test the hypotheses in the above examples at $\alpha = 0.05$ and $\beta = 0.10$, the optimal three-stage design requires $n_1 = 8$ and rejects H_1 if the first eight outcomes are all failures. If $Y_8 > 0$, then $n_2 = 16$ additional patients are treated, and H_1 is rejected if $Y_{24} \leqslant 8$, whereas an additional $n_3 = 39$ patients are treated if $Y_{24} > 8$. The final test has cutoffs (24,25). This design has $E_0(N) = 33.7$, and this slight gain over the analogous Simon optimal two-stage design is typical for small to moderate values of p_0.

Bellisant, Benichou, and Chastang [44] present a simulation study evaluating several multistage phase II designs, including those of Fleming [8] and Herson [7], and designs based on the sequential probability ratio test (SPRT) and the triangular test (TT), and they compare these to the single-stage design. They document the reduction in average sample size obtained by interim monitoring compared to the single-stage approach, as well as the increase in average sample size as the number of patients per stage is increased. Their investigation includes designs based on the SPRT and TT with continuous monitoring, which might be considered hypothesis-test-based alternatives to the Bayesian strategy proposed by Thall and Simon [13,14].

It may be argued that the real purpose of a phase II trial is to obtain a reasonably reliable estimate of the response rate of E, and that interim stopping rules should be imposed mainly to protect patients from inferior treatments. Even for a trial based on tests of hypotheses, a confidence interval for p is of interest at its conclusion. The confidence interval procedure, say at 95%, regards the success probability p as a fixed unknown parameter and the computed interval as a single realization of a random phenomenon that, if it were repeated many times, would contain p between its upper and lower limits 95% of the time. If the success probability is considered to be random rather than fixed, a Bayesian posterior probability interval for the random probability θ_E is appropriate. Alternatively, a frequentist might summarize a Bayesian trial by a confidence interval for the unknown fixed parameter p, with the Bayesian decision rules simply viewed from a frequentist point of view.

When computing a confidence interval, one must account for interim decision rules, since the probability distribution of the confidence interval bounds depends upon the sequences of patient responses and failures possible in the trial. Methods for adjusting confidence intervals computed after trials with interim stopping rules have been discussed by a number of authors, including Jennison and Turnbull [15], Tsiatis, Rosner, and Mehta [16], Atkinson and Brown [17], and Duffy and Santner [18]. For example, if the three-stage Fleming design described above were to run to conclusion with 20 total successes out of $M = 50$ patients, then the correct 95% confidence interval for p that accounts for the interim stopping rules is [0.268–0.556]. If the interim rules are incorrectly ignored, then the corresponding exact Clopper–Pearson [5] confidence interval would be [0.282–0.548]. Continuous monitoring may produce even larger descrepancies. Consider a Bayesian phase II design with only a lower stopping bound, specifically a trial with $M = 42$ that stops if $Y_n/n \leq 0/10$, 1/15, 2/21, 3/27, 4/33, or 5/38. If the trial runs to completion with $Y_{42} = 7$ responses, then the correct 95% confidence interval for p is [0.076–0.360], while the interval that ignores the lower stopping bound is [0.086–0.314].

In contrast, a Bayesian probability interval is based solely on the final data and ignores any interim stopping rules. Based on a noninformative beta (0.4,1.6) prior, i.e., having mean 0.20 and $a + b = 2$, the 95% probability interval running from the 2.5th to 97.5th percentiles of the posterior distribution of θ_E in the above example would be [0.074–0.341]. That is, $\Pr[0.074 < \theta_E < 0.341 \mid Y_{42} = 7] = 0.95$, and this would be the posterior probability interval regardless of the design that produced the final 7/42. The fundamental difference is that the (frequentist) confidence interval for p must be adjusted for interim stopping rules, whereas the Bayesian posterior probability interval for θ_E requires no such adjustment. Also, the unadjusted confidence interval, appropriate following a single-stage design, is nearly identical to the Bayesian probability interval based on a flat prior with $a + b = 2$.

Although multistage and continuous-monitoring designs require a coniderably greater effort in the conduct of the trial, this approach makes use of information ignored by single-stage designs. In particular, continuous monitoring is most protective in the case of a treatment having poor efficacy or an unacceptably high rate of an adverse event. The decision to use a design with continuous monitoring, a multistage design with several interim decisions, a two-stage design with one interim decision, or a single-stage design with a test only at the end should be based in part upon practical considerations and the feasibility of conducting the trial as designed.

An overriding consideration in designing any clinical trial is the logistical aspects of its conduct; hence, the design must provide a balance between scientific goals and what realistically may be implemented in the clinic. A design that either is overly complex or ignores important clinical phenomena is likely to be violated in practice, often out of clinical necessity. The data resulting from such a trial may be unreliable or misleading. A simple example is a cancer chemotherapy trial design that provides rules for monitoring tumor shrinkage but no formal rules for monitoring toxicity. This example is discussed in the next section.

Multiple outcomes

The designs discussed in the preceding sections are based on a single binary outcome. Patient response in clinical trials is an inherently multidimensional phenomenon, however, with the possibility of both adverse events and efficacy outcomes. In addition to evaluating treatment efficacy, a phase II trial must determine whether an experimental treatment is sufficiently safe to allow its evaluation in a large randomized trial. Moreover, responses may occur at two or more stages of the trial, often reflecting the interaction between patient response and subsequent treatment selection in certain clinical settings, as in the case of bone marrow transplantation.

The simplest example is that of a typical cancer chemotherapy trial in which efficacy is evaluated in terms of the usual binary response variable, and acute toxicity is also monitored. If both variables are recorded and toxicity, like response, is scored as a binary variable, then four outcomes are possible. This example illustrates the more general setting in which one efficacy event and one adverse event must be monitored. An important issue in constructing interim stopping rules corresponding to both response and toxicity is the degree of interdependence between these two events, since they are seldom independent. From the viewpoint of safety monitoring, since a high toxicity rate often is associated with a high response rate, a stopping rule that terminates the trial early if the observed toxicity rate is unacceptably high is also likely to terminate a trial of an agent with a high response rate. A common practice in the conduct of phase II trials is to construct the design solely in terms of response, but also to include an informal and often vaguely defined stopping rule for toxicity. This practice results in trials having operating characteristics that are very different from the nominal values obtained from decision rules based on response but ignoring toxicity.

Table 3 presents two hypothetical examples, each having the same marginal probabilities Pr[Response] = 0.40 and Pr[Toxicity] = 0.25, but very different joint probabilities. In case 1, the toxicity rates among responders and nonresponders are very different: 50% of responders suffer toxicity, compared to only 8.3% of nonresponders. This illustrates a double-bind typical of chemotherapy and radiotherapy trials, where decreasing dosage or intensity to reduce toxicity is also likely to reduce the response rate as well, and increasing dosage is likely to increase both the response and toxicity rates. Case 2 illustrates an unlikely scenario in which the toxicity rates of the responders and nonresponders both are identical to the overall rate of 25%, so that the two events are independent. In this unrealistic case, monitoring the two outcomes would not be problematic in that the two monitoring rules could be treated independently. In practice, the nature and degree of interdependency between patient outcomes can be assessed only from historical data, and it is appropriate to use this information in constructing monitoring procedures for multiple events.

Thall, Simon, and Estey [26] present a general Bayesian strategy for monitoring multiple outcomes in single-arm clinical trials. Each patient's response is characterized as a multinomial variable that records the specific combination of events occurring for that patient in the course of the trial, as illustrated by table 3. This includes both adverse events and efficacy outcomes, possibly occurring at different study times. The authors use a Dirichlet-multinomial model to accommodate general discrete multivariate responses, and they provide Bayesian decision criteria for early termination of studies with unacceptably high rates of adverse outcomes or with low

Table 3. Possible outcomes in a trial monitoring both response and toxicity

	Hypothetical probabilities	
Patient response	Case 1	Case 2
A_1 = [Response and No Toxicity]	0.20	0.30
A_2 = [Response and Toxicity]	0.20	0.10
A_3 = [No Response and No Toxicity]	0.55	0.45
A_4 = [No Response and Toxicity]	0.05	0.15

rates of desirable outcomes. Each stopping rule is constructed either to control the rate of an adverse event or to achieve a specified level of improvement of an efficacy event rate for the experimental treatment, compared with that of standard therapy. They avoid explicit specification of costs and a loss function, and evaluate the joint behavior of the multiple decision rules using frequentist criteria. Their approach accommodates a broad range of clinical situations, including settings in which observation of certain endpoints is conditional on the occurrence of earlier events. They illustrate the approach with a variety of single-arm cancer trials, including acute leukemia biochemotherapy trials, bone marrow transplantation trials, and an anti-infection trial.

For a simple application of this method, consider the example given in table 3. Denote the vector of random probabilities of the elementary outcomes by $\boldsymbol{\theta} = (\theta_1,\theta_2,\theta_3)$, with $\theta_4 = 1 - \theta_1 - \theta_2 - \theta_3$, and let $X_n = (X_{n,1},X_{n,2},X_{n,3},X_{n,4})$, the numbers of patients with each combination of outcomes out of n scored. Thus, $X_{n,1} + X_{n,2}$ equals the number of patients who respond, and $X_{n,2} + X_{n,4}$ equals the number of patients who experience toxicity; likewise, $\theta_1 + \theta_2 = \text{Pr[Response]}$ and $\theta_2 + \theta_4 = \text{Pr[Toxicity]}$. The total $X_{n,1} + X_{n,2} + X_{n,3} + X_{n,4} = n$, and $X_n \mid \boldsymbol{\theta}$ is multinomially distributed in n and $\boldsymbol{\theta}$.

Suppose that the event rates in case 1 are obtained from a previous study of 60 patients given 'standard' therapy in which the numbers of patients in the four respective outcome categories were (12,12,33,3). Using these data as the parameters of a Dirichlet prior for the standard-treatment success probability vector $\boldsymbol{\theta}_S$, and using a noninformative prior distribution for the experimental probability vector $\boldsymbol{\theta}_E$, the Thall, Simon and Estey approach might proceed as follows: a 90% posterior probability interval of width 0.20 for Pr[Response], i.e., such that at the end of the trial $\Pr[L < \theta_{E,1} + \theta_{E,2} < U] = 0.90$ with $U - L = 0.20$, requires a maximum of 63 patients. Suppose that a 0.15 increase in the mean response rate is desired, and an increase of at most 0.05 in the toxicity rate is considered an acceptable trade-off for achieving the desired improvement in response rate. The trial is terminated early if

$$\Pr[\theta_{S,1} + \theta_{S,2} + 0.15 < \theta_{E,1} + \theta_{E,2} \mid X_n] \leqslant 0.05$$

or

$$\Pr[\theta_{S,2} + \theta_{S,4} + 0.05 < \theta_{E,2} + \theta_{E,4} \mid X_n] \geqslant 0.90.$$

These determine stopping rules based on the comparison of $X_{n,1} + X_{n,2}$ and $X_{n,2} + X_{n,4}$ to explicit numerical boundaries. The criterion probabilities $p_L = 0.05$ and $p_U = 0.90$ were determined by examining various values of p_L and p_U and selecting those giving desirable operating characteristics. For this design, if true probabilities of response and toxicity are the standard mean values 0.40 and 0.25, then the probability of early termination (PET) is 0.80 and the median sample size is 18. If the toxicity rate is $\geqslant 0.30$, i.e., an increase of 0.05 or more, then PET $\geqslant 0.85$, with a median sample size of at most 18 patients; the PET is larger and the sample size is smaller if the toxicity rate is >0.30. If the response rate is 0.55, the targeted 0.15 improvement, and the toxicity probability is maintained at the null rate of 0.25, then PET $= 0.19$ and the median sample size is the trial maximum of 63.

The general approach of Thall, Simon, and Estey [26] can accommodate considerably more complicated settings; their examples include trials in which the number of elementary patient outcomes varies from three to seven, with as many as four monitoring boundaries running simultaneously. Since the computations necessary to obtain the stopping bounds and operating characteristics are quite complicated, a menu-driven computer program in Splus is available. Simulation of each design takes 5 to 10 seconds on a Solbourne 5/600 computer, so stopping bounds for rather complicated settings may be derived and their properties evaluated rather quickly.

Etzioni and Pepe [25] propose a Bayesian criterion for monitoring two adverse outcomes in a pilot toxicity study, in the case where the occurrence of one event precludes occurrence of the other. The probabilities of the adverse events, θ_1 and θ_2, are considered to be random quantities, and Etzioni and Pepe assume that, given θ_1 and θ_2, the numbers of patients X_1 and X_2 who suffer them are binomially distributed. Etzioni and Pepe define excessive toxicity as the event $A = [\theta_1 > a_1 \text{ or } \theta_2 > a_2]$, where a_1 and a_2 are fixed critical thresholds. For a prior distribution on (θ_1,θ_2), Etzioni and Pepe use the piecewise uniform distribution on the unit square $[0,1] \times [0,1]$, which takes on the values $\{2(1 - a_1a_2)\}^{-1}$ if (θ_1,θ_2) is in A and $\{2a_1a_2\}^{-1}$ if (θ_1,θ_2) is not in A so that, in particular, a priori $\Pr[A] = 1/2$. Their monitoring strategy is to stop the trial if the posterior probability of excessive toxicity exceeds a specified cutoff. For example, if $a_1 = 0.30$, $a_2 = 0.50$, and the cutoff is 0.90, then the trial would be terminated at $n = 4$ patients if either all four patients suffer the first event or three suffer the first event and one suffers the second. Etzioni and Pepe also discuss methods for carrying out frequentist inferences at the end of the trial, including computation of a confidence region for (θ_1,θ_2) and a p-value corresponding to a test of hypothesis.

Discussion

In oncology, nearly any clinical trial that is not a dose-finding study and that does not contain a randomized control group is called a phase II trial. Consequently, the phase II category is quite heterogeneous with regard to objectives and characteristics. Unfortunately, these differences are not always recognized, and statistical designs developed for one type of phase II trial are sometimes inappropriately applied to another type.

Many phase II trials are conducted to evaluate the activity of a new drug against a particular kind of cancer. We have called these phase IIA clinical trials. The main objectives of such trials are to determine whether the drug is active and to obtain a rough estimate of the degree of activity. Often, other drugs have previously been shown to be active. The objective of the trial is not to determine whether the new drug is more active than the other drugs. If the new drug is sufficiently active, the next step may be to combine the new drug with one or more existing drugs to try to identify a regimen that is effective in reducing mortality. The decision of which drugs to include in the combination regimen and whether or not to pursue such an approach depends on several factors. These factors include the level of activity of the new drug, the toxicity profile of the new drug in relation to those of other active drugs, and the levels of activity of available active drugs.

Until recently, most statistical designs developed for phase II clinical trials were applicable primarily to the objectives of phase IIA trials. These include the designs of Gehan [6], Schultz et al. [41], Fleming [8], Simon [11], and Therneau et al. [42]. It is particularly important in this setting that the trial be terminated early if the drug is inactive against the disease, so that patients are not subjected unnecessarily to a toxic agent with no evidence of antitumor activity. Whether or not the trial should continue to a target maximum sample size if the drug has been shown to be active will depend on the clinical setting. Often this is useful for gaining additional experience with the drug in a variety of patients to help plan its incorporation into a combination regimen and to plan subsequent trials. When the drug is in short supply, however, proceeding directly to phase III may be preferable. Also, there is a 'window of opportunity' when it is feasible to conduct a phase III trial of a promising new drug, and prolonging the phase II portion of development may be problematic.

Phase IIB trials have the objective of determining whether a new regimen has a level of antidisease activity that is promising relative to the best available regimens. In dealing with combination regimens, sometimes involving a complex sequential treatment program for the patient, it is not relevant to show that the regimen is 'active'. Moreover, it is generally not feasible to incorporate such complex combinations into other treatment programs. Rather, the focus often is on determining whether the combination regimen under test is sufficiently active, compared to the activity level of best available standard therapy, to warrant a phase III trial. Hence, phase IIB trials are inherently comparative. Usually, however, these comparative aspects are suppressed, or at least not addressed directly. This can have two undesirable effects. The first is that the results with the experimental regimen may appear so promising that a phase III trial is difficult to conduct, since randomization to a control arm appears unethical. The second is that the results are misleadingly promising and a phase III trial is conducted when it is not warranted. There are, of course, other possibilities. In general, we believe that the comparative aspects of phase IIB trials should be addressed directly, that specific control groups should be identified, and that uncertainties arising from the use of nonrandomized control groups of finite size should be quantified. The designs of Thall and Simon [12–14,19], and Thall, Simon, and Estey [26] address this. These designs are, however, quite different from those developed for the simple phase IIA trials.

An alternative to conducting a phase IIB trial is to use a phase III randomized design, allowing one or several experimental regimens, with early termination of a treatment arm if early results with that regimen are sufficiently discouraging. The designs described by Ellenberg, and Eisenberger [45], Thall, Simon, Ellenberg, and Shrager [46], Thall, Simon, and Ellenberg [22,23], Wieand and Therneau [47], Schaid, Wieand, and Therneau [48], and Storer [49] are of this type. It is often difficult to organize a phase III trial of an experimental regimen, however, without some earlier phase II experience with that regimen.

The designs discussed here accommodate a broad range of clinical settings and goals for phase II trials. Some issues remain, however. A major problem in phase II trials is that between-patient variability is typically very large, even given specific entry criteria, while phase II trials are relatively small compared to phase III. Consequently, a phase II trial is not unlikely to have a disproportionate number of patients having relatively poor prognosis within the larger patient group being considered. This in turn is likely to lead to the conclusion that the experimental regimen is not promising as a consequence of the patients' characteristics, rather than due to the effects of the regimen itself. Likewise, a large proportion of good-prognosis patients in the trial, which also is not unlikely, might lead to an overly optimistic conclusion regarding the experimental regimen. A design with interim monitoring rules adjusted for observed patient prognostic variables thus would be highly desirable. Our future research will address this issue.

References

1. Storer BE (1989). Design and analysis of phase I clinical trials. *Biometrics* 45:925–937.
2. O'Quigley J, Pepe MS, Fisher L (1990). Continual reassessment method: A practical design for phase I clinical trials in cancer. *Biometrics* 46:33–48.
3. Simon R (1986). Confidence intervals for reporting results of clinical trials. *Ann Intern Med* 105:429–435.
4. Ghosh BK (1979). A comparison of some approximate confidence intervals for the binomial parameter. *J Am Stat Assoc* 74:894–900.
5. Clopper CJ, Pearson ES (1934). The use of confidence of fiducial limits illustrated in the case of the binomial. *Biometrika* 26:404–413.
6. Gehan EA (1961). The determination of the number of patients required in a follow-up trial of a new chemotherapeutic agent. *J Chron Dis* 13:346–353.
7. Herson J (1979). Predictive probability early termination plans for phase II clinical trials. *Biometrics* 35:775–783.
8. Fleming TR (1982). One sample multiple testing procedure for phase II clinical trials. *Biometrics* 38:143–151.
9. Mehta CR, Cain KC (1984). Charts for the early stopping of pilot studies. *J Clin Oncol* 2:676–682.

10. Chang M, Therneau T, Wieand HS (1987). Designs for group sequential Phase II clinical trials. *Biometrics* 43:865–874.
11. Simon R (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clin Trials* 10:1–10.
12. Thall PF, Simon R (1992). Bayesian design and monitoring of phase II clinical trials. *Proceedings of the XVIth International Biometric Conference, Hamilton, New Zealand*, 205–220.
13. Thall PF, Simon R (1994). Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics* 50:337–349.
14. Thall PF, Simon R (1994). A Bayesian approach to establishing sample size and monitoring criteria for phase II clinical trials. *Controlled Clin Trials* 15:463–481.
15. Jennison C, Turnbull BW (1983). Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials. *Technometrics* 25:49–58.
16. Tsiatis AA, Rosner GL, Mehta CR (1984). Exact confidence intervals following a group sequential test. *Biometrics* 40:797–803.
17. Atkinson EN, Brown BW (1985). Confidence limits for probability of response in multistage clinical trials. *Biometrics* 41:741–744.
18. Duffy DE, Santner TJ (1987). Confidence intervals for a binomial parameter based on multistage tests. *Biometrics* 43:81–93.
19. Thall PF, Simon R (1990). Incorporating historical control data in planning phase II clinical trials. *Stat Med* 9:215–228.
20. Simon R, Wittes RE, Ellenberg SS (1985). Randomized phase II clinical trials. *Cancer Treat Rep* 69:1375–1381.
21. Whitehead J (1986). Sample sizes for phase II and III clinical trials: an integrated approach. *Stat Med* 5:459–464.
22. Thall PF, Simon R, Ellenberg SS (1988). Two stage selection and testing designs for comparative clinical trials. *Biometrika* 75:303–310.
23. Thall PF, Simon R, Ellenberg SS (1989). A two-stage design for choosing among several experimental treatments and a control in clinical trials. *Biometrics* 45:537–547.
24. Strauss N, Simon R (in press). Investigating a sequence of randomized phase II trials to discover promising treatments. *Stat Med.*
25. Etzioni R, Pepe MS (in press). Monitoring of a pilot toxicity study with two adverse outcomes. *Stat Med.*
26. Thall PF, Simon R, Estey EH (1995). Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Stat Med* 14:357–379.
27. Sylvester RJ, Staquet MJ (1977). Decision theory and phase II clinical trials in cancer. *Cancer Treat Rep* 64:519–524.
28. Sylvester RJ, Staquet MJ (1980). Design of phase II clinical trials in cancer using decision theory. *Cancer Treat Rep* 64:519–524.
29. Sylvester RJ (1988). A Bayesian approach to the design of phase II clinical trials. *Biometrics* 44:823–836.
30. Thall PF, Estey EH (1993). A Bayesian strategy for screening cancer treatments prior to phase II clinical evaluation. *Stat Med* 12:1197–1211.
31. Whitehead J (1985). Designing phase II studies in the context of a programme of clinical research. *Biometrics* 41:373–383.
32. Anderson JR, Cain KC, Gelber RD (1983). Analysis of survival by tumor response. *J Clin Oncol* 1:710–719.
33. Freedman LS, Spiegelhalter DJ (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *Statistician* 32:153–160.
34. Freedman LS, Spiegelhalter DJ (1989). Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Controlled Clin Trials* 10:357–367.
35. Spiegelhalter DJ, Freedman LS (1983). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Stat Med* 5:1–13.

36. Spiegelhalter DJ, Freedman LS (1988). Bayesian approaches to clinical trials (with discussion). In *Bayesian Statistics*, JM Bernardo, MH DeGroot, DV Lindley, AFM Smith (eds.). Oxford: Clarendon Press, 453–477.
37. Racine A, Grieve AP, Fluhler H, Smith AFM (1986). Bayesian methods in practice: experiences in the pharmaceutical industry (with discussion). *Appl Stat* 36:93–150.
38. Berry DA (1985). Interim analyses in clinical trials: Classical vs. Bayesian approaches. *Stat Med* 4:521–526.
39. Berry DA (1987). Interim analysis in clinical trials: The role of the likelihood principle. *Am Stat* 41:117–122.
40. Palmer CR (1991). A comparative phase II clinical trials procedure for choosing the best of three treatments. *Stat Med* 10:1327–1340.
41. Schultz JR, Nichol FR, Elfring GL, Weed SD (1973). Multiple stage procedures for drug screening. *Biometrics* 29:293–300.
42. Therneau TM, Wieand HS, Chang SM (1990). Optimal designs for a grouped sequential binomial test. *Biometrics* 46:771–781.
43. Garnsey-Ensign L, Gehan EA, Kamen D, Thall PF (1994). An optimal three-stage design for phase II clinical trials. *Stat Med* 13:1727–1736.
44. Bellisant E, Benichou J, Chastang C (1990). Application of the triangular test to phase II cancer clinical trials. *Stat Med* 9:907–917.
45. Ellenberg SS, Eisenberger MA (1985). An efficient design for phase III studies of combination chemotherapies. *Cancer Treat Rep* 69:1147–1154.
46. Thall PF, Simon R, Ellenberg SS, Shrager R (1988). Optimal two-stage designs for clinical trials with binary response. *Stat Med* 7:571–579.
47. Wieand HS, Therneau TM (1987). A two-stage design for randomized trials with binary outcomes. *Controlled Clin Trials* 8:20–28.
48. Schaid DJ, Wieand S, Therneau TM (1990). Optimal two-stage screening designs for survival comparisons. *Biometrika* 77:507–513.
49. Storer BE (1990). A sequential phase II/III trial for binary outcomes. *Stat Med* 9:229–235.

4. Multivariate failure time data

D.Y. Lin

Introduction

The term *multivariate* usually refers to multiple explanatory variables in clinical literature and to multiple response variables in statistics. In this chapter, the latter interpretation is taken. By multivariate failure time data, we thus mean that each patient may experience several events of clinical interest, or that there exists some natural or artificial clustering of observational units that induces dependence among failure times of the same cluster; we shall refer to the former as multiple events data and the latter as clustered data. Examples of multivariate failure time data include the sequence of tumor recurrences or infection episodes, the developments of physical symptoms or diseases in several organ systems, the experiences of visual loss in the left and right eyes, the onsets of a genetic disease among family members, and the appearances of tumors in littermates exposed to a carcinogen. We describe below three clinical trials and one epidemiologic study that involve multivariate failure times.

Example 1: The Colon Cancer Study. A national intergroup trial was conducted in the 1980s to study the drugs levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma [1,2]. Nine hundred and twenty-nine patients with stage C disease were randomly assigned to observation, levamisole alone, or levamisole combined with fluorouracil. The time to cancer recurrence and the survival time were both considered important outcome measure.

Example 2: The CGD Study. Chronic granulomatous disease (CGD) is a group of inherited rare disorders of the immune function characterized by recurrent pyogenic infections that may lead to death. In order to study the ability of gamma interferon to reduce the rate of infections, a placebo-controlled randomized trial was conducted by the International CGD Cooperative Study Group in the late 1980s. Each patient had the potential to experience multiple infections. By the end of the trial, 30 of 65 placebo patients and 14 of 63 patients on gamma interferon had experienced at least one infection. Of the 30 placebo patients who experienced at least one

Peter F. Thall (ed.), RECENT ADVANCES IN CLINICAL TRIAL DESIGN AND ANALYSIS.
Copyright © 1995 Kluwer Academic Publishers, Boston. All rights reserved. ISBN 978-1-4613-5830-5.

infection, 5 experienced two, 4 others experienced three, and 3 had four or more. Of the 14 gamma interferon patients with at least one infection, 4 experienced two and another had a third event. This study was described at greater length by Fleming and Harrington ([3], pp. 162–163). The data are listed in their appendix D.2.

Example 3: The Diabetic Retinopathy Study. The Diabetic Retinopathy Study was conducted by the National Eye Institute to assess the effectiveness of laser photocoagulation in delaying the onset of blindness in patients with diabetic retinopathy [4]. Seventeen hundred and forty-two patients entered the study between 1972 and 1975. One eye of each patient was randomly selected for photocoagulation, and the other eye was observed without treatment. The patients were followed over several years for the occurrence of blindness in the left and right eyes. One anticipates some dependence between a patient's two eyes.

Example 4: The Schizophrenia Study. Dr. Ann E. Pulver of Johns Hopkins University has been conducting a genetic epidemiologic study of schizophrenia [5]. Four hundred and eighty-seven first-degree relatives (273 males, 214 females) of 93 female schizophrenic probands enrolled in the study. (In human genetics, *proband* means the member of the family that brings a family under study.) The number of relatives of a single proband ranges from 1 to 12. An important question is whether the risk of affective illness (depression or mania or both) in the relatives is associated with the age at onset of schizophrenia of the proband. Here, the times to affective illness are expected to be correlated among relatives of the same proband.

In the above examples, the scientific interests center on the effects of covariates (e.g., treatment) on the risk of failure. For univariate failure time data, i.e., a single failure time variable with independent observations, such effects are studied almost exclusively by the Cox [6] proportional hazards model, which includes the commonly used log-rank test as a special case. The analysis of multivariate failure time data is complicated by the dependence of related failure times. With censoring, this dependence poses a greater statistical challenge than (uncensored) longitudinal data. One useful solution that has gained increasing popularity is the marginal hazard approach originated by Wei, Lin, and Weissfeld [7] and Lee, Wei and Amato [8] (hereafter referred to as WLW and LWA), which formulates the marginal distributions of multivariate failure times with the familiar Cox proportional hazards models while leaving the nature of dependence among related failure times completely unspecified. As in the case of longitudinal data [9], simple estimating equations can be constructed to yield consistent and asymptotically normal estimators for the regression parameters, provided only that the marginal models correctly specified, and robust variance-covariance estimators can be obtained that properly account for the dependence.

The purpose of this chapter is to present an overview of the marginal approach with an emphasis on the designs and analysis of clinical trials. This general methodology is described in the next section. In the Examples section, we provide detailed illustrations with the four real examples cited above. A number of related issues are considered in the Discussion section.

Methods

Univariate failure time data

We first review the basic results for the univariate case. Under the proportional hazards model [6], the hazard function for the failure time T associated with a $p \times 1$ vector of possibly time-varying covariates $Z = (Z_1, \ldots, Z_p)'$ is

$$\lambda(t;Z) = \lambda_0(t)e^{\beta' Z(t)},$$

where β is a $p \times 1$ vector of unknown regression parameters, and $\lambda_0(t)$ is an unspecified baseline hazard function. When T is subject to right-censorship, we observe $X = \min(T,C)$ and $\Delta = I(T \leqslant C)$, where C is the censoring time and $I(\mathscr{A})$ indicates, by the values 1 versus 0, whether or not the event $\mathscr{A}$ occurs. Assume that T and C are independent conditional on Z. Let (X_i, Δ_i, Z_i) $(i = 1, \ldots, n)$ be n independent replicates of (X,Δ,Z). Then the partial likelihood function [10] for β is

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{e^{\beta' Z_i(X_i)}}{\sum_{j=1}^{n} Y_j(X_i) e^{\beta' Z_j(X_i)}} \right\}^{\Delta_i},$$

where $Y_j(t) = I(X_j \geqslant t)$. The corresponding score function $\partial \log L(\beta)/\partial\beta$ equals

$$U(\beta) = \sum_{i=1}^{n} \Delta_i \left\{ Z_i(X_i) - \frac{S^{(1)}(\beta,X_i)}{S^{(0)}(\beta,X_i)} \right\},$$

where $S^{(0)}(\beta,t) = \Sigma_{j=1}^{n} Y_j(t) e^{\beta' Z_j(t)}$ and $S^{(1)}(\beta,t) = \Sigma_{j=1}^{n} Y_j(t) e^{\beta' Z_j(t)} Z_j(t)$. The maximum partial likelihood estimator $\hat{\beta}$ is the solution to $\{U(\beta) = 0\}$. Given $\hat{\beta}$, we estimate the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ by $\hat{\Lambda}_0(t) = \Sigma_{i=1}^{n} I(X_i \leqslant t)\Delta_i / S^{(0)}(\hat{\beta},X_i)$ [11]. The corresponding estimator for the baseline survival function $S_0(t)$ is $\hat{S}_0(t) = e^{-\hat{\Lambda}_0(t)}$.

For large n, the score statistic $U(\beta)$ is approximately p-variate normal with mean 0 and with (estimated) covariance matrix $A(\hat{\beta})$, and $\hat{\beta}$ is approximately p-variate normal with mean β and with (estimated) covariance matrix $A^{-1}(\hat{\beta})$, where

$$A(\beta) = -\frac{\partial^2 \log L(\beta)}{\partial\beta^2} = \sum_{i=1}^{n} \Delta_i \left\{ \frac{S^{(2)}(\beta,X_i)}{S^{(0)}(\beta,X_i)} - \frac{S^{(1)}(\beta,X_i) S^{(1)}(\beta,X_i)'}{S^{(0)}(\beta,X_i)^2} \right\},$$

and $S^{(2)}(\beta,t) = \Sigma_{j=1}^{n} Y_j(t)e^{\beta' Z_j(t)}Z_j(t)Z_j(t)'$ [12]. In addition, the survival function estimator $\hat{S}_0(t)$ is approximately normal with mean $S_0(t)$ and with a variance that can be easily estimated [13]. If the assumed Cox model is incorrect, then the estimator $\hat{\beta}$ is approximately normal with a well-defined mean vector and with covariance matrix $A^{-1}(\hat{\beta})B(\hat{\beta})A^{-1}(\hat{\beta})$ [14], where $B(\hat{\beta}) = \Sigma_{i=1}^{n} W_i(\beta)W_i(\beta)'$ and

$$W_i(\beta) = \Delta_i\left\{Z_i(X_i) - \frac{S^{(1)}(\beta,X_i)}{S^{(0)}(\beta,X_i)}\right\} - \sum_{j=1}^{n} \frac{\Delta_j Y_i(X_j)e^{\beta' Z_i(X_j)}}{S^{(0)}(\beta,X_j)}\left\{Z_i(X_j) - \frac{S^{(1)}(\beta,X_j)}{S^{(0)}(\beta,X_j)}\right\}.$$

If $p = 0$, then the survival function estimator $\hat{S}_0(t)$ is equivalent to the renowned Kaplan–Meier estimator. For testing $\beta = 0$, the nonparametric statistic $U'(0)A^{-1}(0)U(0)$ is known as the logrank statistic, especially when Z is a dichotomous variable. Due to these connections, the Cox regression methodology described above encompasses all the commonly used techniques in survival analysis.

Marginal approach for multivariate failure time data

We now consider the multivariate case. Suppose that there are n units and that each unit can potentially experience K different types of failures. The unit corresponds to the patient in the case of multiple events data and to the cluster for clustered data. Specifically, in examples 1 to 3 above, each patient constitutes a unit, and in example 4 the unit is the proband. In the case of multiple events (e.g., examples 1 and 2), there is generally a clear distinction between different failure types so that the numbering of failure types needs to be consistent across units, whereas for clustered data (e.g., examples 3 and 4), the failure types are indistinguishable, so that the ordering of failure types within a unit is arbitrary. To be more specific, cancer recurrence is very different from death in example 1, whereas a left eye is biologically the same as a right eye in example 3. In the latter case, it would be more precise to say that there are K failures of the same type rather than K different types of failures. To keep our statements concise, however, we will allow ourselves to abuse the language. If there are unequal numbers of failure types among the units, as in example 4, we let K be the maximum number of failure types in a unit.

Let T_{ik} be the time when the kth type of failure occurs on the ith unit, and let C_{ik} be the corresponding censoring time. Define $X_{ik} = \min(T_{ik},C_{ik})$ and $\Delta_{ik} = I(T_{ik} \leqslant C_{ik})$. Also, let $Z_{ik} = (Z_{1ik}, \ldots, Z_{pik})'$ denote the covariate vector for the ith unit with respect to the kth type of failure. For each i, the failure time vector $T_i = (T_{i1}, \ldots, T_{iK})$ and the censoring time vector $C_i = (C_{i1}, \ldots, C_{iK})$ are assumed to be independent conditional on the covariate vector $Z_i = (Z'_{i1}, \ldots, Z'_{iK})$. We further assume that (X_i, C_i, Z_i) $(i = 1, \ldots, n)$ are independent and identically distributed random elements. If T_{ik} or Z_{ik} is missing, we set $C_{ik} = 0$, which ensures that $X_{ik} = 0$ and $\Delta_{ik} = 0$. Naturally, such cases make no contribution to the calculation of the statistics. We require that data are missing completely at random [15].

It is natural to formulate the marginal distribution for each type of failure with a proportional hazards model. Depending on whether the baseline hazard functions are identical or are different among the K types of failures, the hazard function of the ith unit for the kth type of failure is

$$\lambda_k(t;Z_{ik}) = \lambda_0(t)e^{\beta' Z_{ik}(t)}, \tag{1}$$

or

$$\lambda_k(t;Z_{ik}) = \lambda_{0k}(t)e^{\beta' Z_{ik}(t)}, \tag{2}$$

where $\lambda_0(t)$ and $\lambda_{0k}(t)$ $(k = 1, \ldots, K)$ are unspecified baseline hazard functions, and $\beta = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of unknown regression parameters. In the case of multiple events data (e.g., examples 1 and 2), it is generally necessary to allow $\lambda_{0k}(t)$ $(k = 1, \ldots, K)$ to be different, whereas for clustered data (e.g., examples 3 and 4), it is often sufficient to assume a common baseline hazard function. In both models (1) and (2), we take β to be the same among the marginal submodels. This entails no loss of generality, since the assumption can always be achieved by introducing appropriate type-specific covariates, as elaborated in the Examples section below. Note that WLW considered model (2) with type-specific regression parameters, whereas LWA studied model (1).

For the moment, pretend that the observations within the same unit are independent. Then the 'partial likelihood functions' for β are

$$\tilde{L}(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ \frac{e^{\beta' Z_{ik}(X_{ik})}}{\sum_{j=1}^{n} \sum_{l=1}^{K} Y_{jl}(X_{ik}) e^{\beta' Z_{jl}(X_{ik})}} \right\}^{\Delta_{ik}} \tag{3}$$

under model (1) and

$$\tilde{L}(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ \frac{e^{\beta' Z_{ik}(X_{ik})}}{\sum_{j=1}^{n} Y_{jk}(X_{ik}) e^{\beta' Z_{jk}(X_{ik})}} \right\}^{\Delta_{ik}} \tag{4}$$

under model (2), where $Y_{ik}(t) = I(X_{ik} \geqslant t)$. Note that equation (3) is the partial likelihood function for Kn independent observations with a common baseline hazard function, whereas equation (4) is obtained by multiplying the partial likelihood functions for the K marginal submodels. The corresponding 'score functions' are

$$\tilde{U}(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \Delta_{ik} \left\{ Z_{ik}(X_{ik}) - \frac{\tilde{S}^{(1)}(\beta, X_{ik})}{\tilde{S}^{(0)}(\beta, X_{ik})} \right\} \tag{5}$$

and

$$\tilde{U}(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \Delta_{ik} \left\{ Z_{ik}(X_{ik}) - \frac{S_k^{(1)}(\beta, X_{ik})}{S_k^{(0)}(\beta, X_{ik})} \right\}, \tag{6}$$

where $S_k^{(0)}(\beta,t) = \Sigma_{j=1}^{n} Y_{jk}(t) e^{\beta' Z_{jk}(t)}$, $S_k^{(1)}(\beta,t) = \Sigma_{j=1}^{n} Y_{jk}(t) e^{\beta' Z_{jk}(t)} Z_{ik}(t)$ $(k = 1, \ldots, K)$, and $\bar{S}^{(\mathrm{r})}(\beta,t) = \Sigma_{k=1}^{K} S_k^{(r)}(\beta,t)$ $(r = 0,1)$. In both cases, we obtain the unique estimator $\tilde{\beta}$ by solving $\{\tilde{U}(\beta) = 0\}$.

Although observations are generally correlated within the same unit, the estimator $\tilde{\beta}$ can be proven to be consistent for β as long as the marginal models are correctly specified. The derivative matrix $-\partial^2 \log \tilde{L}(\beta)/\partial\beta^2|_{\beta=\tilde{\beta}}$, however, does not provide a valid variance-covariance estimator for $\tilde{U}(\beta)$. As shown in WLW and LWA, by approximating $\tilde{U}(\beta)$ with a sum of n independent and identically distributed random vectors, we can establish the asymptotic normality of $\tilde{U}(\beta)$ and obtain its limiting covariance matrix. Then the asymptotic distribution of $\tilde{\beta}$ follows from the Taylor series expansion. The main results are stated in the following paragraph.

For large n and relatively small K, the statistic $\tilde{U}(\beta)$ is approximately p-variate normal with mean 0 and with (estimated) covariance matrix $\tilde{B}(\tilde{\beta}) = \Sigma_{i=1}^{n} \Sigma_{k=1}^{K} \Sigma_{l=1}^{K} \tilde{W}_{ik}(\tilde{\beta}) \tilde{W}_{il}(\tilde{\beta})'$, where under models (1) and (2), respectively,

$$\tilde{W}_{ik}(\beta) = \Delta_{ik} \left\{ Z_{ik}(X_{ik}) - \frac{\bar{S}^{(1)}(\beta, X_{ik})}{\bar{S}^{(0)}(\beta, X_{ik})} \right\} - \sum_{j=1}^{n} \sum_{l=1}^{K} \frac{\Delta_{jl} Y_{ik}(X_{jl}) e^{\beta' Z_{ik}(X_{jl})}}{\bar{S}^{(0)}(\beta, X_{jl})} \left\{ Z_{ik}(X_{jl}) - \frac{\bar{S}^{(1)}(\beta, X_{jl})}{\bar{S}^{(0)}(\beta, X_{jl})} \right\}$$

and

$$\tilde{W}_{ik}(\beta) = \Delta_{ik} \left\{ Z_{ik}(X_{ik}) - \frac{S_k^{(1)}(\beta, X_{ik})}{S_k^{(0)}(\beta, X_{ik})} \right\} - \sum_{j=1}^{n} \frac{\Delta_{jk} Y_{ik}(X_{jk}) e^{\beta' Z_{ik}(X_{jk})}}{S_k^{(0)}(\beta, X_{jk})} \left\{ Z_{ik}(X_{jk}) - \frac{S_k{}^{(1)}(\beta, X_{jk})}{S_k{}^{(0)}(\beta, X_{jk})} \right\}.$$

Furthermore, the estimator $\tilde{\beta}$ is approximately p-variate normal with mean β and with (estimated) covariance matrix $\tilde{D}(\tilde{\beta}) = \tilde{A}^{-1}(\tilde{\beta}) \tilde{B}(\tilde{\beta}) \tilde{A}^{-1}(\tilde{\beta})$, where

$$\tilde{A}(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \Delta_{ik} \left\{ \frac{\bar{S}^{(2)}(\beta, X_{ik})}{\bar{S}^{(0)}(\beta, X_{ik})} - \frac{\bar{S}^{(1)}(\beta, X_{ik}) \bar{S}^{(1)}(\beta, X_{ik})'}{\bar{S}^{(0)}(\beta, X_{ik})^2} \right\}$$

under model (1) and

$$\tilde{A}(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \Delta_{ik} \left\{ \frac{S_k^{(2)}(\beta, X_{ik})}{S_k^{(0)}(\beta, X_{ik})} - \frac{S_k^{(1)}(\beta, X_{ik}) S_k^{(1)}(\beta, X_{ik})'}{S_k^{(0)}(\beta, X_{ik})^2} \right\}$$

under model (2), $S_k^{(2)}(\beta,t) = \Sigma_{j=1}^{n} Y_{jk}(t) e^{\beta' Z_{jk}(t)} Z_{jk}(t) Z_{jk}(t)'$ $(k = 1, \ldots, K)$ and $\bar{S}^{(2)}(\beta,t) = \Sigma_{k=1}^{K} S_k^{(2)}(\beta,t)$.

Note that $\tilde{A}(\beta) = -\partial^2 \log \tilde{L}(\beta)/\partial\beta^2$. In the case of $\mathrm{K} = 1$, the matrix $\tilde{D}(\tilde{\beta})$ reduces to the Lin–Wei robust variance-covariance estimator given at the end of the previous section. If the marginal models are correctly specified and if the failure times within the same unit are independent, then $\tilde{B}(\tilde{\beta})$ is asymptotically equivalent to $\tilde{A}(\tilde{\beta})$. In the sequel, we refer to $\tilde{A}^{-1}(\tilde{\beta})$ and $\tilde{D}(\tilde{\beta})$ as, respectively, the naive and robust variance-covariance estimators for $\tilde{\beta}$, and call $\tilde{U}'(0)\tilde{A}^{-1}(0)\tilde{U}(0)$ and $\tilde{U}'(0)\tilde{B}^{-1}(0)\tilde{U}(0)$ the naive and robust logrank statistics, respectively. Incidentally, a two-sample robust logrank test was previously studied by Wei and Lachin [16]. It is important to realize that the robust logrank test is always valid, i.e., free of any model assumptions, since the marginal models are guaranteed to hold under $\beta = 0$.

In addition to drawing inferences about individual covariate effects, it is often of interest to test hypotheses involving several components of β. The multivariate general linear hypothesis can be expressed as $H_0 : L\beta = d$, where L is an $r \times p$ matrix of constants and d is an $r \times 1$ vector of constants. The robust Wald statistic for testing H_0 is $(L\tilde{\beta} - d)'\{L\tilde{D}(\tilde{\beta})L'\}^{-1}(L\tilde{\beta} - d)$, which has an approximate χ^2 distribution with r degrees of freedom.

Under the independence working assumption, the Breslow-type estimators for $\Lambda_0(t)$ in models (1) and (2) are, respectively, $\hat{\Lambda}_0(t) = \Sigma_{i=1}^{n} \Sigma_{k=1}^{K} I(X_{ik} \leqslant t)$ $\Delta_{ik}/\bar{S}^{(0)}(\tilde{\beta},X_{ik})$ and $\hat{\Lambda}_{0k}(t) = \Sigma_{i=1}^{n} I(X_{ik} \leqslant t)\Delta_{ik}/S_k^{(0)}(\tilde{\beta},X_{ik})$ $(k = 1, \ldots, K)$. These estimators and the corresponding survival function estimators $e^{-\hat{\Lambda}_0(t)}$ and $e^{-\hat{\Lambda}_{0k}(t)}$ $(k = 1, \ldots, K)$ are approximately unbiased and normally distributed [17].

Simulation results

Monte Carlo simulations were conducted to evaluate the aforementioned inference procedures. Paired failure times with marginal hazard rates $e^{\beta Z_{ik}}$ $(i = 1, \ldots, n;\ k = 1,2)$ were generated from Gumbel's [18] bivariate exponential distribution with correlation coefficient equal to 0.25. Note that only one covariate per failure type was used. Since the failure times were generated with a common baseline hazard function, both models (1) and (2) were true. The covariate values were generated by two different designs. Under the first design, $Z_{i1} = 1$ or 0 with equal probability, and $Z_{i2} = 0$ if $Z_{i1} = 1$ and $Z_{i2} = 1$ if $Z_{i1} = 0$. Under the second design, $Z_{i1} = Z_{i2} = 1$ or 0 with equal probability. Note that the first design corresponds to the matched pairs study and the second design to the group randomization study, in which the group is the randomization unit. Under both designs, the paired failure times were censored independently by a uniform random variable on (0,3), resulting in about 30% censored observations. Table 1 summarizes the results for the combinations of $n = 50$, 100, and 200 and $\beta = 0$ and 5. For each combination, 10,000 data sets were generated. We draw the following conclusions from table 1 and related studies:

1. The bias of $\tilde{\beta}$ is negligible. There is also little bias for the robust standard error estimator, at least for large n. The robust Wald (or logrank) test has proper size, though it may be slightly anticonservative in small and moderate samples. These conclusions hold for both designs under both models (1) and (2).

Table 1. Summary statistics for the simulation studies[a]

			Model (1)						Model (2)					
					SEE		Size/Power				SEE		Size/Power	
Design	β	n	Bias	SSE	Na.	Ro.	Na.	Ro.	Bias	SSE	Na.	Ro.	Na.	Ro.
1	0	50	0.002	0.211	0.245	0.206	0.022	0.055	0.001	0.216	0.252	0.205	0.021	0.060
		100	0.001	0.148	0.172	0.146	0.021	0.056	0.001	0.150	0.175	0.145	0.021	0.057
		200	0.001	0.103	0.121	0.103	0.021	0.051	0.001	0.104	0.122	0.103	0.021	0.053
	0.5	50	0.008	0.209	0.240	0.203	0.575	0.707	0.005	0.215	0.246	0.203	0.541	0.697
		100	0.004	0.147	0.168	0.143	0.891	0.942	0.003	0.148	0.171	0.143	0.878	0.939
		200	0.002	0.103	0.118	0.101	0.997	0.999	0.002	0.104	0.119	0.101	0.996	0.999
2	0	50	0.000	0.285	0.249	0.275	0.081	0.055	0.000	0.287	0.252	0.275	0.079	0.058
		100	−0.002	0.198	0.174	0.194	0.083	0.052	−0.002	0.198	0.175	0.194	0.083	0.053
		200	−0.001	0.139	0.122	0.137	0.085	0.054	−0.001	0.139	0.122	0.137	0.085	0.054
	0.5	50	0.007	0.279	0.243	0.267	0.546	0.473	0.007	0.282	0.246	0.268	0.539	0.471
		100	0.003	0.192	0.170	0.189	0.823	0.762	0.002	0.193	0.171	0.189	0.818	0.759
		200	0.002	0.134	0.119	0.133	0.980	0.968	0.002	0.135	0.119	0.133	0.979	0.966

[a] Bias and SSE are, respectively, the sampling bias and sampling standard error of $\tilde{\beta}$. SEE is the sampling mean of the standard error estimates. Na. and Ro. stand for the naive and robust statistics, respectively. The size and power pertain to the 0.05 nominal significance level.

2. The analysis under model (1) tends to be more efficient than that of model (2), as reflected by the sampling standard error of $\tilde{\beta}$ and by the power of the Wald test. The difference, however, is very small, especially for large n.
3. The naive variance estimator considerably overestimates the true sampling variance under the first design and seriously underestimates the true sampling variance under the second design. Consequently, the naive Wald (or logrank) test has much lower power than the robust test under the first design, and the naive test is not valid under the second design.

Marginal vs. conditional approaches for recurrent events

The choice of time scales for recurrence data needs some discussion. In the marginal approach, T_{ik} is defined as the time from study entry to the kth recurrence for the ith patient ($i = 1, \ldots, n$; $k = 1, \ldots, K$). This time scale, termed *total time*, is particularly appealing when the recurrences are of different natures. In some applications, it is of interest to study the times between consecutive recurrences. i.e., *gap times*. The main difficulty in analyzing gap times is that the patients who have not experienced the kth recurrence have to be excluded from the analysis of the gap times between the kth and (k + 1)th recurrences, which violates the assumption of missing completely at random.

Andersen and Gill [19] and Prentice, Williams, and Peterson [20] (hereafter referred to as AG and PWP) have suggested two alternative approaches to analyzing recurrence data. Under the AG multiplicative intensity model, the risk of a recurrent event for a patient satisfies the usual proportional hazards model, and is unaffected by the patient's earlier events unless terms that capture such dependence are included explicitly in the model as covariates. PWP specified that the hazard function at time t for the kth recurrence of the ith unit, conditional on the entire failure, censoring, and covariate history prior to time t in the unit, takes the form

$$\lambda_{ik}(t) = \lambda_{0k}(t)e^{\beta' Z_{ik}(t)} \tag{7}$$

or

$$\lambda_{ik}(t) = \lambda_{0k}(t - t_{k-1})e^{\beta' Z_{ik}(t)}, \tag{8}$$

where t_{k-1} is the time of the $(k - 1)$th failure $(t_0 = 0)$. Model (7) pertains to total times whereas model (8) uses gap times. The interpretation of the parameters in models (7) and (8) is somewhat awkward because they are conditional on the failure and censoring information. Both the AG and PWP models are analyzed by the partial likelihood principle. As demonstrated by WLW, the AG and PWP procedures are sensitive to misspecification of the dependence structure.

For computational purposes, one may cast the AG and PWP methods within the general framework for the marginal approach described previously. By redefining the risk-set indicators $Y_{ik}(t)$ as $I(X_{i,k-1} < t \leqslant X_{ik})$, instead of $I(X_{ik} \geqslant t)$, with $X_{i0} = 0$ $(i = 1, \ldots, n; k = 1, \ldots, K)$, equations (3) and (4) become the partial likelihood functions for the AG model and model (7) of PWP, respectively. The partial likelihood function for model (8) can be obtained from equation (4) by replacing $Y_{jk}(X_{ik})$ with $Y^*_{jk}(G_{ik})$ and $Z_{jk}(X_{ik})$ with $Z_{jk}(X_{j,k-1} + G_{ik})$, where $G_{ik} = X_{ik} - X_{i,k-1}$ and $Y^*_{jk}(t) = I(G_{jk} \geqslant t)$. In either the case of AG or that of PWP, $\tilde{A}^{-1}(\tilde{\beta})$ is the variance-covariance estimator for the resulting parameter estimator $\tilde{\beta}$.

Which of these three approaches should be used to analyze recurrent events? If one is only interested in the overall rate for recurrences of the same nature, the easiest to use seems to be the AG model (with appropriate time-dependent covariates to capture the dependence), especially when there are only a few second recurrences. If the main interest lies in gap times, then the PWP approach may be used. On the other hand, the marginal approach is the most robust for analyzing total times. It is recommended that each of the three types of models be fit to the same data set, since they provide somewhat different insights.

Monitoring clinical trials

Most clinical trials are monitored periodically for early evidence of treatment difference. In this subsection, we show how to monitor clinical trials with multivariate failure time observations. Much of the material presented here is similar to that given by Lin [21], but several extensions are provided.

Let $\tilde{U}(\beta;t)$ and $\tilde{W}_{ik}(\beta;t)$ $(i = 1, \ldots, n;\ k = 1, \ldots, K)$ be the statistics $\tilde{U}(\beta)$ and $\tilde{W}_{ik}(\beta)$ $(i = 1, \ldots, n;\ k = 1, \ldots, K)$ calculated from the data available at the calendar time t. Suppose that interim analyses are conducted at calendar times $t_1 < t_2 < \ldots < t_M$. Then, under H_0: $\beta = \beta_0$, the pM-dimensional random vector $\{\tilde{U}'(\beta_0;t_1), \ldots, \tilde{U}'(\beta_0;t_M)\}'$ is approximately zero-mean normal, the (estimated) covariance matrix between $\tilde{U}(\beta_0;t)$ and $\tilde{U}(\beta_0;t^\dagger)$ being [21]

$$\tilde{B}(\beta_0;t;t^\dagger) = \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \tilde{W}_{ik}(\beta_0;t)\tilde{W}_{il}(\beta_0;t^\dagger)' \quad (t,t^\dagger = t_1, \ldots, t_M).$$

The foregoing joint distribution provides the basis for constructing various stopping rules.

Suppose for the moment that $\tilde{U}$ is one-dimensional. For an α-level sequential test, we reject H_0 at time t_m if the observed absolute value of the standard normal statistic $\tilde{U}(\beta_0;t_m)/\tilde{B}^{1/2}(\beta_0;t_m,t_m)$ exceeds d_m. The boundary values d_m $(m = 1, \ldots, M)$ are determined recursively by the following equations:

$$\Pr\{|G_1| < d_1, \ldots, |G_{m-1}| < d_{m-1}, |G_m| > d_m\} = \alpha_m,$$
$$m = 1, \ldots, M,$$

where $(\alpha_1, \ldots, \alpha_M)$ are a sequence of exit probabilities such that $\Sigma_{m=1}^{M} \alpha_m = \alpha$, and $(G_1, \ldots, G_m)$ is a zero-mean multivariate normal with covariance matrix $\{\tilde{B}(\beta_0;t,t^\dagger)/(\tilde{B}(\beta_0;t,t)\tilde{B}(\beta_0;t^\dagger,t^\dagger))^{1/2};\ t,t^\dagger = t_1, \ldots, t_m\}$. The above probabilities are evaluated by numerical integration [22] for small M and by simulation for large M. The choice of $(\alpha_1, \ldots, \alpha_M)$ was discussed by Slud and Wei [23] and Lin [21].

It is less straightforward to deal with the case of $p > 1$. The simplest solution is to use a one-dimensional summary number such as the maximal absolute value of or a linear combination of the p (standardized) components of $\tilde{U}(\beta_0;t)$ at each look. Since the joint distribution over the M analyses for the resulting summary statistics follows readily from the known distribution of $\{\tilde{U}'(\beta_0;t_1), \ldots, \tilde{U}'(\beta_0;t_M)\}'$, the boundary values may be obtained in a manner similar to that given in the preceding paragraph. The choice of the linear combination was discussed in great detail by Lin [21].

Software availability

All the methods described in the above subsections on regression models have been implemented in a general FORTRAN program called MULCOX2 [24]. Arbitrary patterns of time-dependent covariates and risk-set indicators are allowed in that program. Dr. Terry Therneau of the Mayo Clinic has also developed some SAS and S macros that serve similar purposes to those of MULCOX2. The sequential methods presented in the previous subsection have been implemented in a FORTRAN program called MULSEQ. All the aforementioned programs are available through StatLib.

Examples

In this section, we apply the techniques described in the last section to the four biomedical studies described in the introduction to this chapter. For comparison, both the naive and robust statistics are presented, though the former are generally inappropriate. All the results reported in this section were obtained from MULCOX2 and MULSEQ (except for the last two columns of table 4).

The colon cancer study

In this trial, 315, 310, and 304 patients with stage C colon cancer received observation, levamisole alone, and levamisole combined with fluorouracil, respectively. Patients were enrolled between March 1984 and October 1987. The study was terminated following an interim analysis in September 1989, when levamisole+fluorouracil was found to be significantly more effective in prolonging survival and reducing the risk of cancer recurrence. By the end of the study, 155 patients in the observation group, 144 in *i*he levamisole alone group, and 103 in the levamisole+fluorouracil group had experienced recurrences, and there had been 114, 109, and 78 deaths in the observation, levamisole alone, and levamisole+fluorouracil groups, respectively. For simplicity, we focus only on the comparison between the observation and levamisole+fluorouracil (Lev+5-FU) groups. Thus, the number of units n is 619 and the number of failure types K is 2. We treat recurrence as the first failure type and death as the second. Since recurrences can only occur before deaths, $\lambda_{01}(t)$ must be different from $\lambda_{02}(t)$.

Let us first consider model (2) with type-specific covariates $Z_{i1} = (R_i,0)'$ and $Z_{i2} = (0,R_i)'$ ($i = 1, \ldots, 619$), where

$$R_i = \begin{cases} 1 & \text{if the } i\text{th patient was on Lev+5-FU,} \\ 0 & \text{if the } i\text{th patient was on observation.} \end{cases}$$

Note that $\beta' Z_{i1} = \beta_1 R_i$ and $\beta_1' Z_{i2} = \beta_2 R_i$ so that β_1 and β_2 pertain to the treatment effects on recurrence and death, respectively. This parameterization illustrates the fact alluded to in the above discussion of the marginal approach for multivariate failure time data that assuming a common β for the K marginal submodels does not preclude the use of typespecific parameters. We are essentially fitting two separate standard Cox models to recurrence and death with the treatment indicator as the single covariate in each model, but formulation (2) permits simultaneous estimation of β_1 and β_2 as well as direct estimation of the correlation between the two estimators. We obtain $\tilde{\beta} = (-0.517, -0.398)'$, with naive and robust standard error estimates of $(0.1273,0.1471)'$ and $(0.1266,0.1475)'$, respectively. The closeness between these two sets of standard error estimates is not surprising, because they are asymptotically equivalent under the current parameterization if the assumed marginal models are correct. The standardized parameter estimates (i.e., estimate/standard error) based on the robust variance estimates are $(-4.08,-2.70)$. The naive and robust variance-covariance estimates for $\tilde{\beta}$ are

$$\tilde{A}^{-1}(\tilde{\beta}) = \begin{bmatrix} 0.0162 & 0 \\ 0 & 0.0216 \end{bmatrix}, \quad \tilde{D}(\tilde{\beta}) = \begin{bmatrix} 0.0160 & 0.0144 \\ 0.0144 & 0.0218 \end{bmatrix}.$$

Because of the high correlation between $\tilde{\beta}_1$ and $\tilde{\beta}_2$, the naive and robust tests for a multivariate hypothesis (involving both β_1 and β_2) can be quite different. For example, the logrank statistic for testing $\beta = 0$ is 24.27 using $\tilde{A}(\tilde{\beta})$ and 17.23 using $\tilde{B}(\tilde{\beta})$. The robust Wald statistic for testing $\beta_1 = \beta_2$ is 1.57 whereas the naive test statistic is 0.37. Apparently, there is no convincing evidence for different sizes of treatment effects on cancer recurrence and death.

We now suppose that $\beta_1 = \beta_2 = \beta$. Note that the null hypothesis of no treatment effect on either recurrence or death corresponds to $\beta_1 = \beta_2 = \beta = 0$. As long as β_1 and β_2 are not too far apart, the estimator of β provides a useful summary of the overall treatment difference. By letting $Z_{i1} = Z_{i2} = R_i$, which implies that $\beta' Z_{i1} = \beta' Z_{i2} = \beta R_i$, we obtain $\tilde{\beta} = -0.466$ with naive standard error estimate of 0.096 and robust standard error estimate of 0.128, the corresponding standardized parameter estimates being -4.84 and -3.65, respectively. Unlike the estimation of *separate* treatment effects discussed in the preceding paragraph, the naive and robust standard error estimators for the *common* treatment effect are not asymptotically equivalent if the two failure types are correlated. The use of the robust standardized estimate or the robust log rank statistic (the latter being 13.54) for the common parameter β would enable one to make a single probability statement regarding the overall benefit of Lev+5-FU.

There were some imbalances between the observation and Lev+5-FU groups with respect to certain prognostic factors. Thus, it is desirable to run a confirmatory analysis that adjusts for the prognostic variables. To this end, we fit model (2) with $Z_{i1} = Z_{i2} = (R_i, S_i, D_i, N_i)'$, where

$$S_i = \begin{cases} 1 & \text{if the surgery for the } i\text{th patient took place} \leq 20 \text{ days prior to randomization,} \\ 0 & \text{if the surgery for the } i\text{th patient took place} > 20 \text{ days prior to randomization;} \end{cases}$$

$$D_i = \begin{cases} 1 & \text{if the depth of invasion for the } i\text{th patient was submucosa or muscular layer,} \\ 0 & \text{if the depth of invasion for the } i\text{th patient was serosa;} \end{cases}$$

$$N_i = \begin{cases} 1 & \text{if the number of nodes involved in the } i\text{th patient was 1–4,} \\ 0 & \text{if the number of nodes involved in the } i\text{th patient} > 4. \end{cases}$$

This analysis yields -0.483 as the estimate for the common treatment effect with robust standard error estimate of 0.131. The corresponding standardized parameter estimate is −3.69, so the treatment effect remains significant after adjusting for the prognostic factors. The depth of invasion and the number of nodes are both highly significant.

Eleven of the 619 patients in the observation and Lev+5-FU groups died without cancer recurrences. In the above analyses, recurrence times on those patients were censored at deaths. Strictly speaking, the assumption of conditional independence between the failure time and the censoring time may not be completely satisfied in such cases. To avoid this problem, one may consider deaths without recurrences as events for the first failure type. Then the first failure time variable is interpreted as recurrence-free survival time, i.e., time to either cancer recurrence or death, whichever occurs first. For this study, very similar results were obtained between the two approaches, mainly because less than 2% of the patients died without recurrences, compared to 42% who had recurrences first. For the model considered in the preceding paragraph, i.e., model (2) with $Z_{i1} = Z_{i2} = (R_i,S_i,D_i,N_i)'$, we obtain $\tilde{\beta}_1 = -0.467$ with robust standard error estimate of 0.130 when deaths without recurrences are treated as events for the first failure type.

In this study, four interim analyses were planned, but the study was terminated at the second analysis. The formal stopping rule was defined on mortality, though recurrence was also taken into consideration in the decision making. Here we demonstrate how the approach described above in the section on monitoring clinical trials might have been employed to construct a formal stopping rule based on both recurrence and death. Let $\tilde{U}_1(\beta;t)$ and $\tilde{U}_2(\beta;t)$ be,, respectively, the components of the score function for recurrence and death calculated at calendar time t, and let $\tilde{B}_{jl}(\beta_0;t,t^\dagger)$ be the covariance between $\tilde{U}_j(\beta_0;t)$ and $\tilde{U}_l(\beta_0;t^\dagger)$ under H_0: $\beta = \beta_0$. Table 2 displays the observed values of $\tilde{U}_1(0;t)$ and $\tilde{U}_2(0;t)$ along with the variance-covariance estimates for the first two interim looks. Assume that $\alpha_1 = \alpha_2 = 0.005$, $\alpha_3 = 0.01$, and $\alpha_4 = 0.03$. Let R(t) be the sum of the two *standardized* score statistics at time t divided by its variance. Note that the sum of the *non-standardized* $\tilde{U}_1(0;t)$ and $\tilde{U}_2(0;t)$ is equivalent to the one-dimensional score

Table 2. Observed score statistics and variance-covariance estimates at the first two interim looks for the Colon Cancer Study

				$\tilde{B}_{jl}\ (0;t,t^\dagger)$			
				$t^\dagger = t_1$		$t^\dagger = t_2$	
t	j	$\tilde{U}_j\ (0;t)$	$\tilde{U}_j\ (0;t)/\tilde{B}_{jj}^{1/2}\ (0;t,t)$	$l = 1$	$l = 2$	$l = 1$	$l = 2$
t_1	1	−23.646	−3.842	37.876	16.383	37.732	29.866
t_1	2	−5.099	−1.169		19.025	16.434	19.465
t_2	1	−32.868	−4.097			64.353	42.704
t_2	2	−18.870	−2.725				47.949

statistic $\tilde{U}(0;t)$ for testing H_0: $\beta = 0$, where β is the common regression parameter discussed in the 3rd paragraph of this subsection. From table 2, we obtain $R(t_1) = -2.792$ and $R(t_2) = -3.627$ with 0.764 as the estimated covariance or correlation. By numerical integration, the boundary values associated with the specified α_1 and α_2 and the correlation estimate are $d_1 = 2.807$ and $d_2 = 2.721$. Thus, $|R(t_1)| < d_1$ and $|R(t_2)| > d_2$. For further illustration, let $Q(t)$ be the maximum of the absolute values of the two standardized score statistics at time t. The covariance matrix for the four standardized statistics is easily obtained from table 2. Given this covariance matrix and the aforementioned α_1 and α_2, the boundary values for the two maxima are found to be $d_1 = 3.006$ and $d_2 = 2.915$. Clearly, both $Q(t_1)$ and $Q(t_2)$ exceed the boundary values. Thus, one would have terminated the trial at the first look because of the early evidence for the benefit of Lev+5-FU in reducing the risk of cancer recurrence.

The CGD study

The main statistical analysis of the CGD study was based on time to the first infection. By fitting the standard Cox model with the treatment indicator R ($R_i = 1$ if the ith patient was on gamma interferon and $R_i = 0$ otherwise) as the single covariate, we obtain $\hat{\beta} = -1.094$ with standard error estimate of 0.335. Note that our numbers are different from those of Fleming and Harrington ([8], p. 163) because they used only the infections that had occurred by the interim analysis cutoff, whereas we make use of the additional data on occurrence of infections between the interim analysis cutoff and the final study visit for each patient. Appendix D.2 of Fleming and Harrington [3] contains the full data set used here.

Since the investigators were interested in how gamma interferon reduces the *rate* of infections, it seems desirable to incorporate into the analysis the additional data on recurrent events. The simplest way is to fit the AG multiplicative intensity model for all infection episodes with R as the single covariate. Under this Markov model, the estimate of treatment effect is -1.097 with an estimated standard of 0.261. This analysis assumes that the patient's risk for a new infection at a given time is not altered by the pattern of prior infections. As an attempt to accommodate the dependence of infection patterns, we add to the preceding model a time-dependent covariate, which indicates by the value 1 versus 0 whether or not the patient had an infection within the previous 60 days. The parameter estimate for this covariate is 0.712 with standard error estimate of 0.293, which is highly significant. In this semiMarkov model, the estimate for the treatment parameter becomes -0.989 with standard error estimate of 0.266.

The representation of the infection history by simple time-dependent covariates may be inadequate. To avoid specifying the nature of dependence, we use the marginal approach. In this application, T_{ik} is the time from study enrollment to the kth infection for the ith patient, and C_{ik} is the time from study enrollment to the final study visit for the ith patient. Since there were very few fourth infections, we will study only the first three infections. To estimate separate treatment effects for the three failure types, we fit model (2) with $Z_{i1} = (R_i,0,0)'$, $Z_{i2} = (0,R_i,0)'$, and $Z_{i3} = (0,0,R_i)'$ $(i = 1, \ldots, 128)$; to estimate an overall treatment effect, we fit model (2) with $Z_{ik} = R_i$ $(i = 1, \ldots, 128; k = 1,2,3)$. The results of these analyses are summarized in table 3. For comparison, we also display the results for the PWP approach using the same covariates as the marginal approach, as well as those of the two AG models. Note that the last column in table 3 pertains to a common regression parameter. Note also that the results for the AG models reported in the last paragraph were based on *all* infections, whereas those of table 3 are restricted to the *first three* infections only.

As shown in table 3, using any of the three approaches, one arrives at the conclusion that gamma interferon indeed reduces the infection rate substantially. Compared to the PWP and AG methods, the marginal approach gives a somewhat larger estimate of the common treatment parameter along with a larger standard error estimate. Note that, for testing no overall treatment benefit, the marginal approach is always valid, whereas the validity of PWP and AG methods depends on correct specification of the dependence structure. It is interesting to observe that the PWP approach does not yield significant treatment effects for the second and third infections.

The Diabetic Retinopathy Study

We confine our attention to a subset of the data from the Diabetic Retinopathy Study (DRS) that was previously analyzed by Huster, Brookmeyer, and Self [25] and Liang, Self, and Chang [26]. The analysis subset is a 50% sample of the high-risk patients as defined by DRS criteria ($n = 197$). By the

Table 3. Estimates of treatment effects for the CGD Study[a]

	Infection number			
Methods	1	2	3	1~3
Marginal	−1.094	−1.231	−2.063	−1.215
	(0.335)	(0.538)	(1.019)	(0.353)
PWP				
Total time	−1.094	0.151	−1.279	−0.859
	(0.335)	(0.566)	(1.084)	(0.280)
Gap time	−1.094	−0.090	−1.077	−0.872
	(0.335)	(0.537)	(1.084)	(0.279)
AG				
Markov	—	—	—	−1.020
	—	—	—	(0.267)
Semi-Markov	—	—	—	−0.943
	—	—	—	(0.269)

[a] The standard error estimates are given in parentheses.

end of the study, 54 treated eyes and 101 control eyes in this subsample had developed blindness.

In this example, each patient could potentially experience blindness in both eyes; therefore, there are two failure types, with $k = 1$ and 2 denoting the left and right eyes, respectively. Since there are no biological differences between the left and right eyes, it is natural to assume a common baseline hazard function for the two failure types.

As mentioned in the introduction to this chapter, the main hypothesis of interest is whether laser photocoagulation delays the occurrence of blindness. Because juvenile and adult diabetes have very different courses, it is desirable to examine how the age at onset of diabetes may affect the time to blindness. Following Huster et al. and Liang et al., we consider model (1) with $Z_{ik} = (Z_{1ik}, Z_{2ik}, Z_{3ik})'$ $(i1, \ldots, 197;\ k = 1,2)$, where

$$Z_{1ik} = \begin{cases} 1 & \text{if the } k\text{th eye of the } i\text{th patient was on treatment,} \\ 0 & \text{otherwise;} \end{cases}$$

$$Z_{2ik} = \begin{cases} 1 & \text{if the } i\text{th patient had adult onset diabetes,} \\ 0 & \text{if the } i\text{th patient had juvenile onset diabetes;} \end{cases}$$

and $Z_{3ik} = Z_{1ik} * Z_{2ik}$. The results of our analysis are presented in table 4 along with those of Huster et al. and Liang et al.

The robust standard error estimates are appreciably smaller than the naive estimates. The treatment appears to be effective, and this effect is much stronger for adult onset diabetes than for juvenile diabetes. The Liang et al. estimating function is similar to our equation (5), but they replaced $\bar{S}^{(1)}/\bar{S}^{(0)}$ by an analogue that exploits pairwise comparisons of independent observations. Their method produced very similar parameter estimates to ours, and their standard error estimates are almost identical to our robust ones. Huster et al. specified a Weibull baseline hazard function for model (1). Their parameter estimates are fairly close to $\tilde{\beta}$, whereas their standard error estimates are similar to be naive estimates.

Table 4. Estimates of regression parameters for the Diabetic Retinopathy Study[a]

	Methods			
Covariate	Naive	Robust	Liang	Huster
Treatment (Z_1)	−0.425	−0.425	−0.422	−0.43
	(0.218)	(0.185)	(0.185)	(0.22)
Diabetic type (Z_2)	0.341	0.341	0.340	0.37
	(0.199)	(0.196)	(0.196)	(0.20)
Interaction ($Z_1 * Z_2$)	−0.846	−0.846	−0.844	−0.84
	(0.351)	(0.304)	(0.303)	(0.35)

[a] The standard error estimates are given in parentheses.

The Schizophrenia Study

In this ongoing genetic epidemiologic study, the failure time is the age at diagnosis of affective illness for the relative. There are only 31 events out of the 487 relatives in the current database. The covariate of major interest, the proband's age, has been dichotomized at 16 years. The gender of the relative is also expected to be predictive. We assume that gender is the only characteristic that differentiates relatives of the same proband. It is then natural to consider model (1) with $Z_{ik} = (Z_{1ik}, Z_{2ik})'$ ($i = 1, \ldots, 93$; $k = 1, \ldots, 12$), where

$$Z_{1ik} = \begin{cases} 1 & \text{if the age at onset of the } i\text{th proband} \leqslant 16, \\ 0 & \text{otherwise;} \end{cases}$$

$$Z_{2ik} = \begin{cases} 1 & \text{if the } k\text{th relative of the } i\text{th proband is male,} \\ 0 & \text{if the } k\text{th relative of the } i\text{th proband is female.} \end{cases}$$

Note that we set $K = 12$, the maximal number of relatives for a proband. Since the ordering of relatives within a family is arbitrary, we suppose that the missing components occupy the tail portion of the failure vector $T_i = (T_{i1}, \ldots, T_{i,12})'$ for each i. We obtain $\tilde{\beta} = (-0.238, -1.244)'$ with naive and robust standard error estimates of $(0.489, 0.411)'$ and $(0.517, 0.408)'$, respectively. Therefore, the proband's age at onset is not significant, whereas the relative's gender is. The failure to establish an association between the familial risk and the proband's age at onset may be due to the small number of events. We intend to reanalyze the data after further follow-up.

Discussion

The estimating functions (5) and (6) were derived under the independence working assumption. As in the case of longitudinal data [9], it may be more efficient to use estimating functions that take into account the nature of dependence explicitly. This amounts to forming certain linear combinations of the contributions to functions (5) or (6) from the K types of failures. The resulting estimators remain consistent and asymptotically normal with estimable covariance matrices under mild regularity conditions on the weight matrices. Because of the censoring and the nonlinear nature of the Cox model, however, it is difficult to construct optimal weight matrices. In her Ph.D. dissertation, Cai [27] suggested the use of the inverse matrix of the covariance functions between counting process martingales [28] for model (2). Her simulations, however, indicated that the efficiency improvements of the resulting estimators are small unless the correlations of failure times are unusually high. For estimating a common regression parameter, WLW used a linear combination of type-specific parameter estimators that achieved the smallest asymptotic variance among all linear combinations.

For the CGD study, the WLW method estimates the overall treatment effect at -1.103 with standard error estimate of 0.333. In a similar spirit, Lin [21] proposed a weighted sum of the marginal logrank statistics which maximizes asymptotic power against certain local alternatives. Further research into dependent working models is warranted.

It is important to assess the adequacy of the marginal models. A simple method for examining the key proportional hazards assumption is to test for the significance of interaction terms between covariates and t or $\log t$, as was originally suggested by Cox [6]. One may also compare parameter estimators with different weight matrices [29]. Recently, elaborate techniques for checking (univariate) survival models have been developed by using martingale-based residuals [30–32]. Generalizations of these methods to the multivariate setting are currently being investigated by C. Spiekerman in his University of Washington Ph.D. dissertation.

When designing clinical trials, one is often faced with the task of sample size calculation. Due to the complicated nature of the robust variance, it is difficult to derive a simple sample-size formula. The easiest solution is to use the formula for the independence case [33] and then adjust the sample size upward or downward depending on whether it is a clustered or matched study. A more precise approach is simulation. Under the latter approach, one would, as in the above subsection on simulation results, specify the joint distribution of the multivariate failure times along with the usual design parameters and then obtain the empirical powers of the resulting robust logrank test for various sample sizes.

In many applications, failure times are broadly grouped. Most commonly, they arise when the (continuous) failure time is subject to interval grouping. In other instances, the time measurement may be truly discrete, as, for example, when the time represents the number of attempts required to successfully perform a certain task. For the univariate failure time variable, Prentice and Gloeckler [34] studied a grouped data version of the Cox proportional hazards model. Recently, Guo and Lin [35] extended the work of Prentice and Gloeckler to the multivariate setting. Their procedures are essentially the discrete versions of those described above in the section about multivariate failure time data.

A useful alternative to the proportional hazards model is the accelerated failure time model, which relates the logarithm of the failure time linearly to the covariates. Semiparametric inference for this model has received considerable attention in the last few years [36,37]. Recently, Lin and Wei [38] and Lee, Wei, and Ying [39] applied the ideas of WLW and LWA, respectively, to the case of accelerated failure time models.

As mentioned above, a few patients in the colon cancer study died without cancer recurrences. In most of the analyses reported here, the recurrence times of those patients were censored at their death times. Strictly speaking, the resulting relative risk estimators pertain to the so-called cause-specific hazard function ([40], p. 167) rather than the usual net hazard function. Pepe and Mori [41] discussed the limitations of cause-specific hazard functions and advocated the use of cumulative incidence functions and conditional probabilities. Recently, Lin, Robins, and Wei [42] proposed a bivariate accelerated failure time model for the times to cancer recurrence and death for the two treatment groups and constructed semiparametric procedures for comparing the two marginal distributions of recurrence.

The marginal approach exploited in this chapter treats the dependence of related failure times as a nuisance. In contrast, a number of authors [43–46] have studied the so-called frailty models, which explicitly formulate the nature of dependence. To be specific, the hazard function for the ith unit with respect to the kth type of failure, given the frailty Q_i, takes the form

$$\lambda_{ik}(t;Z_{ik},Q_i) = Q_i\lambda_0(t)e^{\beta' Z_{ik}(t)}, \tag{9}$$

where the frailty variables Q_i ($i = 1, \ldots, n$) are postulated to follow a given parametric distribution. Conditional on Q_i ($i = 1, \ldots, n$), the failure times are assumed to be independent. Note that β in equation (9) generally needs to be interpreted conditionally on the unobservable frailty. There has been considerable controversy over whether the unconditional specification of the marginal hazard approach or the conditional specification of the frailty model approach is more naturally related to the underlying mechanisms. The latter approach is expected to be more efficient than the former, provided that the frailty distribution is correctly specified. However, the types of dependence encompassed by frailty models are quite limited, and the model fitting is rather cumbersome. So far there has not been a general large-sample theory for frailty models, though significant progress is being made. The interested reader is referred to the recent text of Andersen et al. [19] for an excellent exposition of frailty models.

Acknowledgments

The author is grateful to Professors Thomas Fleming, David Harrington, Steven Self, and Kung-Yee Liang for providing the data files of the Colon Cancer, CGD, DRS, and Schizophrenia Studies, respectively. He would also like to thank Professors L.J. Wei and Zhiliang Ying for many stimulating discussions on the subject over the years. This research was supported by the National Institutes of Health Grants R29 GM-47845 and R01 AI-291968.

References

1. Moertel CG, Fleming TR, McDonald JS, et al. (1990). Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *N Engl J Med* 322:352–358.
2. Fleming TR (1992). Evaluating therapeutic interventions: some issues and experiences (with discussion). *Stat Sci* 7:428–456.

3. Fleming TR, Harrington DP (1991). *Counting Processes and Survival Analysis*. New York: Wiley.
4. Diabetic Retinopathy Study Research Group (1981). Diabetic retinopathy study. *Invest Ophthalmol Visual Sci* 21 (Part 2):149–226.
5. Pulver AE, Liang K-Y (1991). Estimating effects of proband characteristics on familial risk: II. The association between age at onset and familial risk in the Maryland schizophrenia sample. *Genet Epidemiol* 8:339–350.
6. Cox DR (1972). Regression models and life-tables (with discussion). *J R Stat Soc B* 34:187–220.
7. Wei LJ, Lin DY, Weissfeld L (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J Am Stat Assoc* 84:1065–1073.
8. Lee EW, Wei LJ, Amato DA (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, JP Klein, PK Goel (eds.). Dordrecht: Kluwer Academic Publishers, 237–247.
9. Liang K-Y, Zeger SL (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73:13–22.
10. Cox DR (1975). Partial likelihood. *Biometrika* 62:269–276.
11. Breslow N (1972). Contribution to the discussion of the paper by DR Cox, *J R Stat Soc B* 34:216–217.
12. Andersen PK, Gill RD (1982). Cox's regression model for counting processes: a large sample study. *Ann Stat* 10:1100–1120.
13. Lin DY, Fleming TR, Wei LJ (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika* 81:73–81.
14. Lin DY, Wei LJ (1989). The robust inference for the Cox proportional hazards model. *J Am Stat Assoc* 84:1074–1078.
15. Rubin DB (1976). Inference and missing values. *Biometrika* 63:81–92.
16. Wei LJ, Lachin JM (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *J Am Stat Assoc* 79:653–661.
17. Spiekerman CF, Lin DY (submitted). Survival function estimation for correlated failure time data under the marginal Cox model.
18. Gumbel, EJ (1960). Bivariate exponential distributions. *J Am Stat Assoc* 55:698–707.
19. Andersen PK, Borgan ϕ, Gill RD, Keiding N (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.
20. Prentice RL, Williams BJ, Peterson AV (1981). On the regression analysis of multivariate failure time data. *Biometrika* 68:373–379.
21. Lin DY (1991). Nonparametric sequential testing in clinical trials with incomplete multivariate observations. *Biometrika* 78:123–131.
22. Schervish MJ (1984). Multivariate normal probabilities with error bound. *Appl Statist* 33:81–94. Corrections 34:103–104.
23. Slud EV, Wei LJ (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *J Am Stat Assoc* 77:862–868.
24. Lin DY (1993). MULCOX2: a general computer program for the Cox regression analysis of multivariate failure time data. *Comput Methods Programs Biomed* 40:279–293.
25. Huster WJ, Brookmeyer R, Self SG (1989). Modelling paired survival data with covariates. *Biometrics* 45:145–156.
26. Liang K-Y, Self SG, Chang Y-C (1993). Modelling marginal hazards in multivariate failure time data. *J R Stat Soc B* 55:441–453.
27. Cai J (1992). *Generalized Estimating Equations for Censored Multivariate Failure Time Data*. Unpublished Ph.D. dissertation, Departrent. of Biostatistics, University of Washington, Seattle, WA.
28. Prentice RL, Cai J (1992). Covariance and survival function estimation using censored multivariate failure time data. *Biometrika* 79:495–512.
29. Lin DY (1991). Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators. *J Am Stat Assoc* 86:725–728.

30. Barlow WE, Prentice RL (1988). Residuals for relative risk regression. *Biometrika* 75:65–74.
31. Therneau TM, Grambsch PM, Fleming TR (1990). Martingale-based residuals for survival models. *Biometrika* 77:147–160.
32. Lin DY, Wei LJ, Ying Z (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* 80:573–581.
33. Schoenfeld DA (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics* 34:57–67.
34. Prentice RL, Gloeckler LA (1978). Regression analysis of grouped survival data with applications to breast cancer data. *Biometrics* 34:57–67.
35. Guo SW, Lin DY (1994). Regression analysis of multivariate grouped survival data. *Biometrics* 50:632–639.
36. Tsiatis AA (1990). Estimating regression parameters using linear rank tests for censored data. *Ann Stat* 18:354–372.
37. Wei LJ, Ying Z, Lin DY (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika* 77:845–851.
38. Lin JS, Wei LJ (1992). Linear regression analysis for multivariate failure time observations. *J Am Stat Assoc* 87:1071–1097.
39. Lee EW, Wei LJ, Ying Z (1993). Linear regression analysis for highly stratified failure time data. *J Am Stat Assoc* 88:557–565.
40. Kalbfleisch JD, Prentice RL (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.
41. Pepe MS, and Mori M (1993). Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Stat Med* 12:737–751.
42. Lin DY, Robins JM, Wei LJ (1995). Comparing two failure time distributions in the presence of informative censoring. *Biometrika*, in press.
43. Clayton D, Cuzick J (1985). Multivariate generalizations of the proportional hazards model. *J R Stat Soc A* 148 (Part 2):82–117.
44. Hougaard P (1987). Modelling multivariate survival. *Scand J Stat* 14:291–304.
45. Oakes D (1989). Bivariate survival models induced by frailties. *J Am Stat Assoc* 84:487–493.
46. Nielsen GG, Gill RD, Andersen PK, Sϕrensen TIA (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scand J Stat* 19:25–43.

5. Goodness-of-fit and diagnostics for proportional hazards regression models

Patricia M. Grambsch

Introduction

A common clinical study design follows patients over time, recording endpoint events as they occur for each individual. In a cancer clinical trial with death as the endpoint, there can be at most one event per patient. In other cases, multiple events are possible — for example, studies of recurrent infections in bone marrow transplantation recipients. The study goal is to model the event rate as a function of covariates measured at baseline. In a clinical trial, these would typically include the treatment group, measures of disease severity, patient age, and other sociodemographic variables. The proportional hazards regression model is a popular tool.

In counting process notation [1] the data consist of n independent triples

$$\{N_i(t), Y_i(t), Z_i; \quad i = 1, \ldots, n, \quad t \in [0,\tau]\}$$

observed over the time period $[0,\tau]$. The counting process $N_i(t)$ is the number of events observed for subject i in the interval $[0,t]$. The predictable process $Y_i(t)$ is a 1–0 indicator process showing whether or not subject i is at risk and under observation at time t. Its sample paths are left-continuous step functions. Z_i is a p-vector of covariates measured on subject i at baseline. Let $\{F_t, t \geqslant 0\}$ be the right-continuous filtration, specifying the process history:

$$F_t = \sigma\{Z_i(u), N_i(u), Y_i(u+): \quad 0 \leqslant u \leqslant t, \quad i = 1, \ldots, n\}.$$

The proportional hazards regression model assumes that the intensity process (with respect to the process history) for individual i can be written

$$h_i(t)dt = Y_i(t)\exp\left\{\sum_{j=1}^{p} \beta_j f_j(Z_{ij})\right\}\lambda_0(t)dt. \tag{1}$$

The f_j's are known functions of the covariates and the β_j's are parameters to be estimated. The function $\lambda_0(t)$, the baseline intensity function, is an unspecified function of time to be estimated. The intensity process can be interpreted as the probability of an event in the next brief time period, dt. In

Peter F. Thall (ed.), RECENT ADVANCES IN CLINICAL TRIAL DESIGN AND ANALYSIS.

ISBN 978-1-4613-5830-5.

the case of survival data, $\exp\{\Sigma_{j=1}^{p}\beta_j f_j(Z_{ij})\}\lambda_0(t)$ is the hazard function. The expected number of events for individual i can be found by integrating the intensity process:

$$E_i = EN_i(\tau) = \int_0^\tau E\{Y_i(s)\}\lambda_0(s)\exp\{\Sigma\beta_j f_j(Z_{ij})\}ds. \tag{2}$$

There are two basic assumptions for this model. The first is the assumption of proportional hazards. This means that the hazard ratio $h_i(t)/h_j(t)$ for any two individuals i and j does not depend on time. The second assumption is that of the functional form of the covariates. The model assumes that the impact on the log hazard for the jth covariate is linear in $f_j(Z_j)$. There are numerous statistical techniques for assessing these two assumptions. Good reviews can be found in Andersen et al. ([2], chapter VII.3), Kay [3], and Lin and Wei [4]. It is not our purpose to provide another encyclopedic review. Rather, we will focus on a few simple graphical techniques that we have found useful in indicating departures from the assumptions. These techniques are based on residuals computed after fitting a proportional hazards regression model to the data. We will use suitably scaled martingale residuals to assess the functional form for covariates, and suitably scaled Schoenfeld residuals to detect departures from proportional hazards.

Martingale and Schoenfeld residuals

Inference for the proportional hazards model typically proceeds by maximizing the log partial likelihood to estimate β [5]. Let f_i be the p-vector with jth element $f_i(Z_{ij})$. The log partial likelihood is

$$L(\beta) = \sum_{i=1}^{n}\int_0^\tau\left[Y_i(t)\beta' f_i - \log\left\{\sum_l Y_l(t)e^{\beta' f_l}\right\}\right]dN_i(t). \tag{3}$$

Define

$$S^{(r)}(\beta,t) = \sum_{i=1}^{n} Y_i(t)\exp\{\beta' f_i\}f_i^{\otimes r}$$

for $r = 0, 1, 2$ where, for a column vector a, $a^{\otimes 2}$ denotes the outer product aa', $a^{\otimes 1}$ denotes the vector a and $a^{\otimes 0}$ denotes the scalar 1. The conditional weighted mean and variance of the covariate vector at time t are

$$\bar{f}(\beta,t) = S^{(1)}(\beta,t)/S^{(0)}(\beta,t)$$

and

$$V(\beta,t) = \frac{S^{(2)}(\beta,t)}{S^{(0)}(\beta,t)} - \left\{\frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}\right\}^{\otimes 2}.$$

The score vector process $\frac{\partial L(\beta)}{\partial \beta}$ is

$$U(\beta,t) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ f_i - \bar{f}(\beta,s) \right\} dN_i(t), \tag{4}$$

and the maximum partial likelihood estimator $\hat{\beta}$ is the solution to the estimating equation $U(\beta,\tau) = 0$. Under suitable regularity conditions [6], it has the usual properties of a maximum likelihood estimator. It is consistent and asymptotically normally distributed with mean equal to the true value of β and variance/covariance matrix consistently estimated by the inverse of the observed information matrix,

$$\mathscr{I}(\hat{\beta}) = \sum_{i=1}^{n} \int_0^{\tau} V(\hat{\beta},s) dN_i(s). \tag{5}$$

The cumulative baseline intensity function $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ is estimated by [7]

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\Sigma dN_i(s)}{\Sigma Y_i(s) \exp\{\hat{\beta}' f_i\}}.$$

The martingale residuals [8] are motivated by counting process martingales. If the assumed model is correct, then

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\beta' f_i} \lambda_0(s) ds \quad (i = 1, \dots, n) \tag{6}$$

is a subject-specific martingale. The martingale residual is defined as

$$\hat{M}_i(t) = N_i(t) - \int_0^t Y_i(s) e^{\hat{\beta}' f_i} d\hat{\Lambda}_0(s), \tag{7}$$

with M_i as shorthand for $\hat{M}_i(\tau)$. The residual can be interpreted as the observed number of events minus the conditional expected number, given the at-risk process, for the time period $[0,t]$. These residuals have many of the properties of residuals from normal theory linear models:

$$\Sigma M_i(t) = 0 \quad \text{for any } t,$$

and asymptotically

$$E(\hat{M}_i) = \text{Cov}(\hat{M}_i, \hat{M}_j) = 0.$$

The partial residuals introduced by Schoenfeld [9] are the increments in the score process (equation (4)). Suppose the d events in the study occur at ordered times $0 < t_1 < t_2 < \dots < t_k < \dots < t_d < \tau$. Let $f_{(k)}$ be the covariate vector of the subject with an event at time t_k. Let $r_k(\beta) = f_{(k)} - \bar{f}(\beta,t_k)$ and $V_k(\beta) = V(\beta,t_k)$. Then the Schoenfeld residuals are given by $\hat{r}_k = r_k(\hat{\beta})$, $k = 1, \dots, d$. This residual can be interpreted as the difference between the covariate vector observed to have an event at t and its conditional expectation, given those at risk at that time. If the assumed model is correct, then the $r_k(\beta)$'s are uncorrelated with mean 0 and conditional variance matrix $V(\beta,t_k)$ given $\mathscr{F}t_k-$. Clearly, $\Sigma_k\hat{r}_k = 0$. The Schoenfeld residuals, $\hat{r}_1, \ldots, \hat{r}_d$, have means that are asymptotically 0 and $\operatorname{cov}(\hat{r}_l,\hat{r}_m)$, which may be consistently estimated by $\delta_{lm}\hat{V}_l - \hat{V}_l\mathscr{I}(\hat{\beta})^{-1}\hat{V}_m$ where $\hat{V}_l = V(\hat{\beta},t_l)$ [9]. Note that $\mathscr{I}(\hat{\beta}) = \Sigma_l\hat{V}_l$.

Assessing proportionality

Suppose the proportional hazards assumption is not met. A simple way to conceptualize this departure is to allow one or more covariates to have a time-varying effect on the intensity process, thus forcing hazard ratios to vary with time. An example is a treatment effect decreasing over time. Let $\beta_j(t) \equiv \beta_j + \theta_j g_j(t)$; $j = 1, \ldots, p$, where each $g_j(t)$ is a predictable process with respect to the history filtration. Informally, this means that its behavior at t is determined by its behavior on $(0,t)$ for any t. Examples include continuous deterministic functions of time and left-continuous stochastic processes that are functions of the counting processes up to $t-$. For identifiability, we assume that g varies about 0. The intensity process becomes

$$h_i(t)dt = Y_i(t)\exp[\{\beta + G(t)\theta\}'f_i]\,\lambda_0(t)dt, \tag{8}$$

where $G(t)$ is a $p \times p$ diagonal matrix with $G_{jj}(t) = g_j(t)$ and θ is a p-vector with jth element θ_j. The model follows proportional hazards when $\theta = 0$; otherwise, it has time-varying coefficients.

We generalize Schoenfeld's approach [9], which considered a time-varying coefficient for only one covariate. Using his Taylor's expansion argument, we get

$$E(\hat{r}_k|\mathscr{F}_{t_k-}) \approx V(\beta,t_k)G_k\theta,$$

where $G_k = G(t_k)$. Further, if $|\theta|$ is not too large,

$$\operatorname{cov}(\hat{r}_l,\hat{r}_m) \approx \delta_{lm}V_l - V_l(\Sigma V_k)^{-1}V_m$$

for large samples. Let $r_k^* = \hat{V}_k^{-1}\hat{r}_k$ be the scaled Schoenfeld residual. Then

$$E(r_k^*) \approx E(G_k)\theta, \tag{9}$$

$$\operatorname{var}(r_k^*) \approx V_k^{-1} - \mathscr{I}(\beta)^{-1}, \tag{10}$$

$$\operatorname{cov}(r_k^*,r_l^*) \approx -\mathscr{I}(\beta)^{-1}. \tag{11}$$

Equations (9) to (11) suggest a standard linear model for r_k^*. Generalized least squares gives

$$\hat{\theta} = D^{-1}\Sigma G_k\hat{r}_k, \tag{12}$$

with

$$D = \Sigma G_k\hat{V}_kG_k' - (\Sigma G_k\hat{V}_k)(\Sigma\hat{V}_k)^{-1}(\Sigma G_k\hat{V}_k)'. \tag{13}$$

Under $\mathscr{H}_0$, the asymptotic variance of $n^{-1/2}\Sigma G_k\hat{r}_k$ can be consistently estimated by $n^{-1}D$, leading to an asymptotic χ^2 test statistic on p degrees of freedom:

$$T(G) = (\Sigma G_k\hat{r}_k)'D^{-1}(\Sigma G_k\hat{r}_k). \tag{14}$$

For any single variate, this derivation shows that on average, r_k^* is proportional to $g(t_k)$, a rescaling of the time axis, suggesting a plot of r_k^* versus $g(t_k)$. The test is essentially a test for a nonzero slope in a generalized linear regression of the scaled residuals on the chosen rescaling of time.

These results, motivated by least-squares heuristics, can be formalized. The estimator (12) is a one-step Newton–Raphson estimator of θ, starting from the estimated proportional hazards model with β at $\hat{\beta}$ and $\theta = 0$. The test statistic is the Rao score test of proportionality ($\mathscr{H}_0$: $\theta = 0$) based on the partial likelihood. A proof of the asymptotic distribution of $T(G)$ under $\mathscr{H}_0$ follows from standard results for score processes using counting process theory. The fully iterated maximum partial likelihood estimator of θ_j could be obtained by additional Newton–Raphson steps. The connection between score tests and generalized least squares has been observed in many settings. It is useful to note that it holds in this less standard framework as well.

Many of the standard tests for proportionality are, in fact, $T(G)$ tests for particular choices of G. Typically, G is diagonal, so we will refer to a univariate function $g(t)$. When $g(t)$ is a user-specified function of time, $T(G)$ is the score test for the addition of the time-dependent variable $g(t)f(Z)$ to the model, the test suggested by Cox [5]. Letting g be piecewise constant on nonoverlapping time intervals (with the interval boundaries and constants prespecified by the investigator) gives the score test proposed by O'Quigley and Pessione [10], which generalizes tests proposed by Schoenfeld [11] and Moreau, O'Quigley, and Mesbah [12]. If $g(t) = \Sigma N_i(t-)$, one obtains essentially the test proposed by Harrell [13], who tested the correlation between the Schoenfeld residuals and the rank of the event times. Lin [14] suggested comparing $\hat{\beta}$ to the solution $\hat{\beta}_g$ of the weighted estimating equation

$$\Sigma G_k r_k(\beta_g) = 0$$

with $g(t)$ one of the scalar weight functions commonly chosen for weighted logrank tests, such as the left-continuous version of the Kaplan–Meier estimator. He showed that asymptotically $\hat{\beta}-\hat{\beta}_g$ is multivariate normal with mean 0 and a variance matrix derived from martingale counting process theory. If the estimator $\hat{\beta}_g$ were based on a one-step Newton–Raphson algorithm starting from $\hat{\beta}$, his test would be identical to $T(G)$. Finally, let $g_j(t_1) = 0$ and $g_j(t_{k+1}) = a_j^2\hat{r}_{jk}$, where $j = 1, \ldots, p$. This gives the test statistic of Nagelkerke, Oosting, and Hart [15], who suggested using the serial correlation of the Schoenfeld residuals for a univariate predictor, or, for multivariate covariates, the correlation of a weighted sum, $a'\hat{r}_k$. They proposed $a = \hat{\beta}$ as a natural choice for the weights, followed by examination of individual covariates if the test is significant.

The $T(G)$ tests provide a general regression framework for tests of proportionality. In turn, they are themselves a special case of more general weighted score tests. The numerator of the test, $\Sigma G_k \hat{r}_k$, can be written as

$$\int_0^{\tau} \sum_{i=1}^{n} Y_i(s) G(s) \, [f_i - \bar{f}(\hat{\beta},s)] \, dN_i(s),$$

which is a multivariate version of the nonparametric test of Jones and Crowley [16,17]. They show that many of the commonly used tests in survival analysis, including the Harrington and Fleming [18] family of tests for comparing two groups, the Tarone and Ware family of tests for comparing k groups, the s-sample trend test of Tarone and Ware [19], and the logit rank test [20] have this form.

These statistical tests can be supplemented by graphical diagnostics. Schoenfeld [9] and Lin [14] proposed plotting the components of the Schoenfeld residuals against event times. Pettitt and Bin Daud [21] plotted the Schoenfeld residual components scaled by the corresponding inverse diagonal elements of the $\hat{V}_k$'s. Redefine the scaled Schoenfeld residuals as $\hat{\beta} + r_k^*$. Equations (8) and (9) imply that the p componentwise plots of scaled residuals versus t_k with a scatterplot smooth superimposed to estimate the mean will suggest the functional form of $\beta(t)$. The mean scaled residual is a one-step Taylor approximation to the functional form.

Many scatterplot smoothers are readily available. Typically, they are linear smoothers. For a scatterplot of two n-vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$, a linear smoother gives the fitted smooth as $\hat{\boldsymbol{y}} = \boldsymbol{L}\boldsymbol{y}$, where the $n \times n$ smoothing matrix $\boldsymbol{L}$ depends on $\boldsymbol{x}$, the smoothing bandwidth, and possibly user-specified weights, but not on $\boldsymbol{y}$. Examples include kernel smooths, regression splines, smoothing splines and locally weighted regression (e.g., loess [22,23,24,25]).

The use of a smoother introduces an unavoidable element of subjectivity, since the user must choose the type of smoother and the bandwidth. The type of smoother is usually not very important, so long as the smoother is sensitive to local rather than global features of the data set ([26], chapters 3 and 4).

Computer-intensive algorithms based on cross-validation can be used to optimize the choice of smoothing bandwidth. A simpler approach is to use the equivalent degrees of freedom of the smooth, as defined by Hastie and Tibshirani [26]. By analogy with linear regression, one can use $\mathrm{tr}(\boldsymbol{L}\boldsymbol{L}') - 1$ as the degrees of freedom. The matrix $\boldsymbol{L}$ corresponds to the projection matrix of multiple linear regression and, in the case of regression including an intercept term, $\mathrm{tr}(\boldsymbol{L}\boldsymbol{L}') - 1$ gives the 'degrees of freedom for regression' familiar from regression ANOVA tables. The equivalent degrees of freedom term is inversely proportional to the bandwidth of the smooth. The analogy to linear regression can often suggest a reasonable amount of smoothing for any particular data set. When one expects fairly smoothly time-varying coefficients and has a moderate number of events (at least 30), smoothers with 3 to 5 degrees of freedom are often adequate, we have found.

Another computational simplification is frequently possible when plotting scaled Schoenfeld residuals. The computation of $\hat{V}_k$ at each event time may not be necessary; for most data sets, the covariance matrix of the at-risk covariates is fairly stable until the last few events. One can substitute the average value $\bar{V} = \mathscr{I}(\hat{\beta})/d$. This estimate may even be preferable, because the last $\hat{V}_k$'s are based on only a few subjects each and may even be singular. A refinement is to consider $\mathscr{I}(\hat{\beta})/d^*$, where d^* is the number of events where $\hat{V}_k$ is invertible. The computations for $\bar{V}$ and the r_k's are unchanged if there are tied events, but the usual caveats about biased estimates apply if the proportion of ties is large.

We illustrate the performance of these techniques on some data sets. The first example [12] comes from a clinical trial comparing chemotherapy alone to chemotherapy plus radiation for patients with locally advanced nonresectable gastric carcinoma, using death as the endpoint. A proportional hazards model with treatment as a binary covariate (chemotherapy with radiation = 0, chemotherapy alone = 1) gave $\hat{\beta} = -0.267$ with standard error = 0.23. The test of Moreau, O'Quigley, and Mesbah [12] led to a rejection of proportional hazards. Figure 1 shows the scaled Schoenfeld residuals plotted against death times. As is typical for data with a single binary covariate, the residuals fall into two bands. Positive values mark death times for chemotherapy patients, and negative values mark chemotherapy plus radiation deaths. Two three-degrees-of-freedom scatterplot smooths are superimposed, namely, linear gaussian loess (solid line) and a cubic natural spline (dotted line) with two knots at the tertiles of the death times. Stablein, Carter, and Novak [27] fit a time-varying coefficient model to these data:

$$h_i(t) = \lambda_0(t)\exp\left[\left\{\beta_0 + \theta_1\left(\frac{t}{30}\right) + \theta_2\left(\frac{t}{30}\right)^2\right\}Z_i\right].$$

They found $\hat{\beta}(t) = -1.866 + 0.1768(t/30) - 0.0028(t/30)^2$, and this estimate is also shown on the plot (dashed line). The three curves are quite similar. All suggest that the beneficial effect of radiation on relative risk decreases over time and has effectively disappeared by about one year.

Approximate confidence intervals for the smooths can be computed under the null hypothesis of proportional hazards and can be used to assess any departures from proportional hazards observed on the plots. Let $\hat{Y} = L\hat{R}^* + \hat{\beta}$, where $\boldsymbol{L}$ is the $d \times d$ smoothing matrix and $\hat{R}^*$ has kth row $\hat{r}_k^* = \hat{V}_k{}^{-1}\hat{r}_k$. Under the null hypothesis of proportional hazards, each column j of $\hat{R}^*$ is asymptotically multivariate normal with mean 0 and variance–covariance S_j, say. Note that conditioning on the event times and risk sets, $\text{var}(\hat{r}_k^*)$, can be estimated consistently by $\hat{V}_k^{-1} - \mathscr{I}^{-1}$ and $\text{cov}(\hat{r}_k^*, \hat{r}_l^*)$ by $-\mathscr{I}^{-1}$. Therefore, we estimate S_j by $A_j - \mathscr{I}_{j,j}^{-1}J$, where A_j is a $d \times d$ diagonal matrix whose kth diagonal element is $\hat{V}_{k,j,j}^{-1}$, J is a $d \times d$ matrix of 1's. The variance of $\hat{Y}$ involves the variance of $\hat{\beta}_j$. Due to the optimality of the partial likelihood score equation under $\mathscr{H}_0$ [28], $\hat{\beta}$ and the Schoenfeld

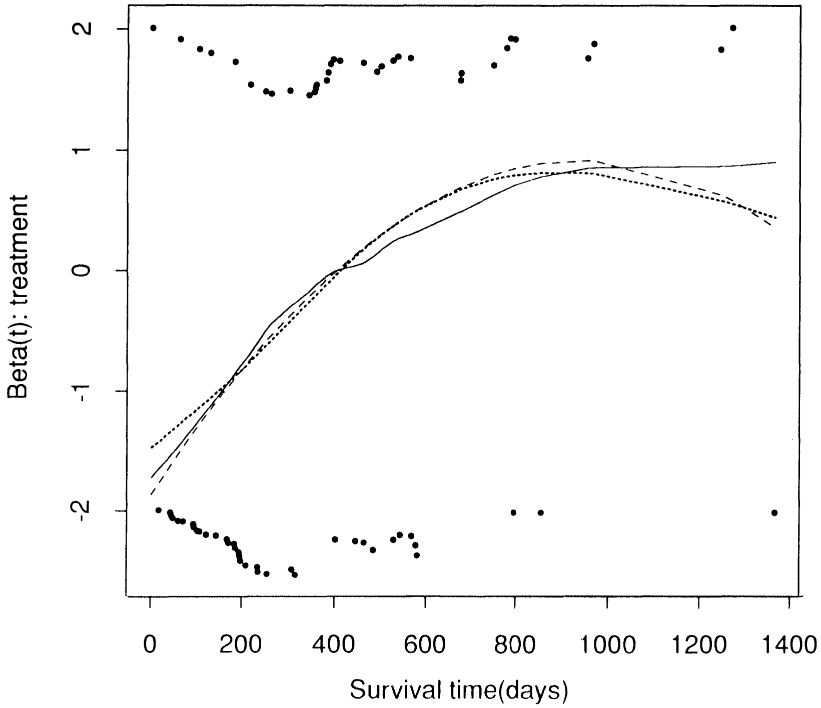


Figure 1. Scaled Schoenfeld residuals $+\hat{\beta}$. A loess smooth, three degrees of freedom (solid line), a natural spline smooth, three degrees of freedom (dotted line), and the quadratic time × covariate interaction model of Stablein, Carter, and Novak (dashed line) are superimposed.

residuals are asymptotically uncorrelated. Conditioning on the observed failure times so that we can treat L as deterministic, the jth column of $\hat{Y}$ is asymptotically normal with mean $\beta_j 1$ and variance matrix LA_jL'. Confidence intervals can be formed by standard linear model calculation, e.g., Scheffé intervals using the rank of LS_jL' for simultaneous confidence bands or z-intervals for pointwise estimates.

As a simplification, one might consider using $\bar{V} = d^{-1}\mathcal{I}$ in place of $\hat{V}_k$. Then S_j becomes $\mathcal{I}_{jj}^{-1}\{(d + 1)I - J\}$. For smoothers based on linear regression against a basis matrix X, such as splines, the calculation is very similar to that for the ordinary regression hat matrix $H = X(X'X)^{-1}X'$, in that a $d \times d$ matrix need not ever be explicitly constructed. The problem is computationally more complex for the loess smoother [23]. However, one typically wants pointwise confidence intervals at only a few time points, and therefore only the rows of L corresponding to those points are needed. As a practical matter, we have rarely found the simplified approach to lead to conclusions different from those obtained using the $\hat{V}_k$'s. Therneau [29] has written a set of S [30] or S-Plus [31] functions that will implement these diagnostics.

The second example comes from a study comparing stage II and stage IIA ovarian cancer patients on time from treatment initiation to disease progression [32]. There were 35 patients (15 stage II and 20 stage IIA) with 22 events. Gill and Schumacher's test [33], a version of Lin's test for a scalar covariate, showed that one could reject the proportional hazards assumption for these data. Figure 2 shows the scaled Schoenfeld residuals from a proportional hazards model with stage as a binary covariate (stage II = 0 and stage IIA = 1). The two methods for scaling, event-specific variance $V_k^{-1}\hat{r}_k$ (right panel) and average variance $\mathcal{I}^{-1}\hat{r}_k/d^*$ (left panel), give similar residuals. The loess scatterplot smoother with three equivalent degrees of freedom was used. It was modified for event-specific variance scaling to include weights, with each point weighted inversely by its estimated variance under $\mathcal{H}_0$, $\hat{V}_k^{-1} - \mathcal{I}^{-1}$. Superimposed on each plot are 90% pointwise confidence intervals at $t = 34$, 199, 270, 370, and 451 days. The smooths and confidence intervals for the two scaling techniques are virtually identical and suggest that the risk of progression for stage IIA relative to stage II increases over time, particularly after 200 days. Simulation results for other models, including multiple covariates, are given by Grambsch and Therneau [34].

Assessing functional form

The intensity process

$$h_i(t)dt = Y_i(t)\exp\left\{\sum_{j=1}^{p}\beta_j f_j(Z_{ij})\right\}\lambda_0(t)dt$$

requires specification of the $f_j(Z)$'s, the functional forms for the covariates. In the case of quantitative covariates, standard practice involves using the identity function and assuming that the intensity is log-linear in Z_j. However, the true functional form may involve Z_j^2, $\ln Z_j$, $I_{\{Z_j>c\}}$, or some other transform. This issue was examined by Therneau, Grambsch, and Fleming [35]. They suggested assessing covariates one at a time. To examine any particular covariate, one would fit a proportional hazards model omitting that covariate and compute the martingale residuals scaled by dividing by the proportion of failures in the data set. A plot of the scaled martingale residuals against the covariate of interest with a scatterplot smooth superimposed would reveal the functional form for that covariate. However, as the authors pointed out, this plot did not work well when the covariate effects were large. A careful reading of their proof shows that it also requires that the covariate of interest be uncorrelated with other covariates in the model.

A refinement of their diagnostic plot can be motivated from two directions. The first is the close relationship between counting process models and Poisson regression. Poisson regression can be viewed as a special case of the generalized linear model (GLM), and the GLM partial residual plot for

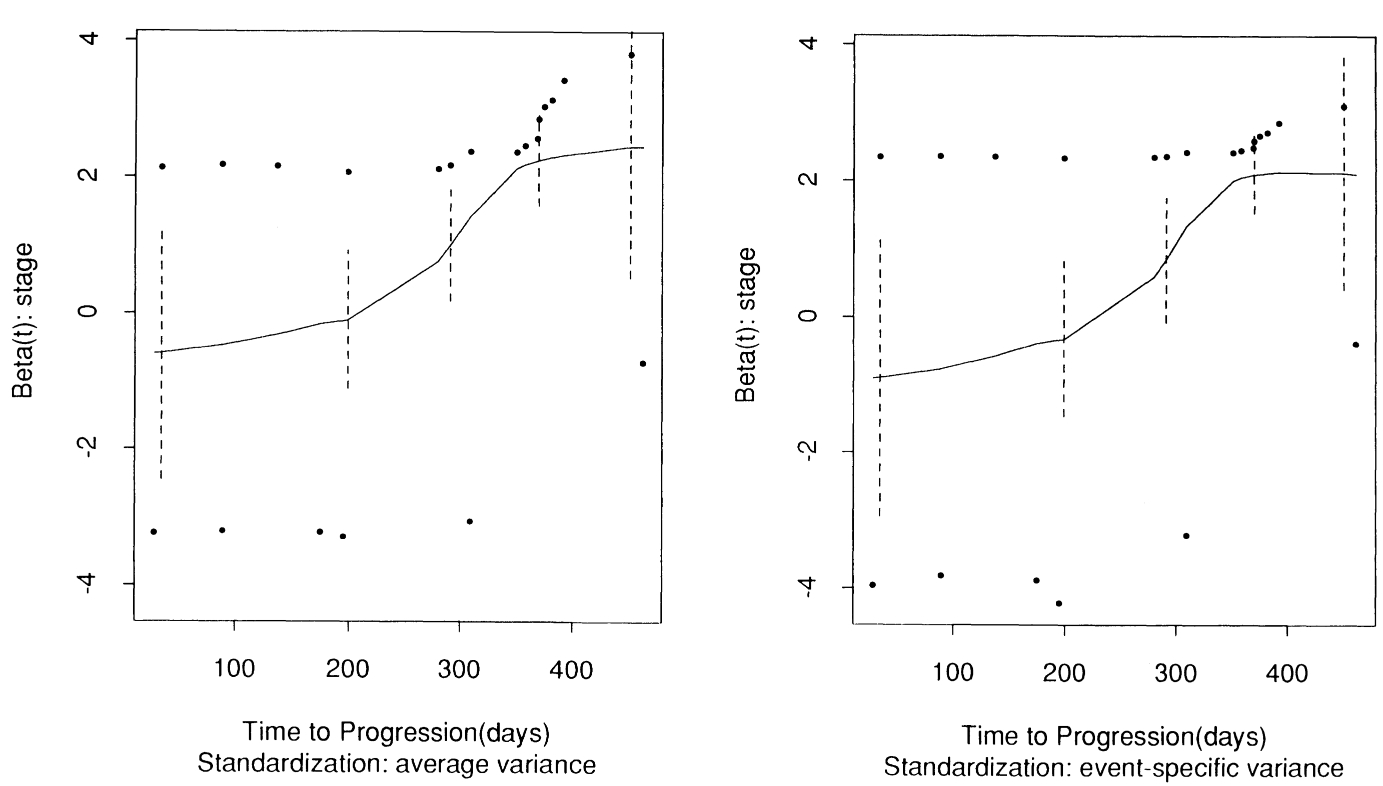


Figure 2. Scaled Schoenfeld residuals $+\hat{\beta}$ with a loess smooth, three equivalent degrees of freedom, superimposed. The left panel uses average variance standardization. The right panel uses event-specific variance scaling with a weighted loess smooth with weights inversely proportional to the variance. Ninety percent pointwise confidence intervals are shown at 34, 191, 291, 370, and 451 days.

assessing functional form in Poisson regression ([36], p. 402) suggests a modification of the martingale plot for counting process data. Let X denote the covariate of interest. GLM partial residuals are computed after fitting a model that includes X. This requires an initial guess for the functional form for X. If one expects a monotonic relationship between X and the intensity, a log-linear form is often adequate, and we use that form for illustration. The investigator fits the model:

$$h_i(t)dt = Y_i(t)\exp\left\{\sum_{j=1}^{p-1} \beta_j f_j(Z_{ij}) + \gamma X_i\right\}\lambda_0(t)dt$$

Let

$$\hat{E}_i = \int_0^\tau Y_i(s)\exp\left\{\sum_{j=1}^{p-1} \hat{\beta}_j f_j(Z_{ij}) + \hat{\gamma} X_i\right\}\hat{\lambda}_0(t)dt$$

denote the expected count for the ith individual and let $\hat{M}_i = N_i - \hat{E}_i$ be the martingale residual. Then the GLIM partial residual is

$$\frac{\hat{M}_i}{\hat{E}_i} + \hat{\gamma} X_i.$$

McCullagh and Nelder [36] recommend plotting the partial residual against X as an informal check for the correctness of the initial guess: 'The partial residual plot, if smoothed, can be remarkably informative even for binary data.'

For counting process data, it is important to use a weighted scatterplot smooth with $\hat{E}_i$'s as weights. This procedure corresponds to weighting inversely as the variance because the variance of M_i is the expected value of E_i. More importantly, the weights offer a bias correction. Even if the initial guess as to functional form is correct, an unweighted approach can have serious bias. As an instructive example, consider the case where there are a large number of replicates at each unique value of X. Let the replicates be indexed by the additional subscript j, $j = 1, \ldots, n_i$, at each unique value X_i. A simple smooth would just take the mean of the replicates at each X_i. The weighted smooth involves the weighted average of $\hat{M}_{ij}/\hat{E}_{ij}$ with weights $\hat{E}_{ij}$, which is

$$\frac{\Sigma_j N_{ij}/n_i}{\Sigma_j \hat{E}_{ij}/n_i} - 1.$$

The unweighted smooth is

$$\left(\sum_j \frac{N_{ij}}{\hat{E}_{ij}}\right)\frac{1}{n_i}.$$

The smoothed partial residual then adds $\hat{\gamma} X_i$ to each mean. Let the number of replicates increase. If the postulated model is correct, the weighted smooth will converge to γX_i, since $\Sigma_j N_{ij}/n_i$ and $\Sigma_j \hat{E}_{ij}/n_i$ both converge to

$E(E_i)$. The correct functional form will be seen. The unweighted smooth may be poorly behaved, since $n_i^{-1}\Sigma_j N_{ij}/\hat{E}_{ij}$ need not converge to 1. It is easy to construct examples where it becomes unbounded.

If the postulated functional form for X is not correct, but the other covariates are scaled correctly, a Taylor expansion argument shows that the weighted smoothed partial residual plot can suggest the true functional form, provided the smoother is reasonably unbiased for the true functional form and the initial guess is not too poor (see Grambsch, Therneau, and Fleming [37] for more details).

A Monte Carlo experiment was done to compare the smoothed weighted GLM partial residuals to smoothed unweighted GLM partial residuals and to the original proposal of Therneau, Grambsch, and Fleming [35]. The experiment involved censored survival data with hazard rate $\lambda_0 \exp\{f_1(Z_1) + f_2(Z_2)\}$ with $f_1(Z_1) = 1.7Z_1$ and $f_2(Z_2) = 0.63(Z_2 - 0.6)^3$. The covariates were bivariate normal with common mean 1.5, common standard deviation 0.3, and two different correlations: 0.0 and 0.90. We set $\lambda_0 = \exp\{-f_1(\mu) - f_2(\mu)\}$. Censoring was independent of failure time and uniform on $[0,c]$, with c chosen to give a censoring rate of roughly 27%. One thousand data sets of 133 observations were generated for each correlation. To compute GLM partial residuals, a Cox model with hazard rate log-linear in Z_1 and Z_2 was fit to each data set. For the martingale residuals, a Cox model with hazard rate log-linear in Z_1 but ignoring Z_2 was fit. In each case, the functional form was correct for Z_1 but not for Z_2.

The smoother was a linear gaussian loess with four degrees of freedom. Because the Cox model is semparametric, the intercept in a diagnostic plot is not identifiable. Therefore, each smooth was adjusted by vertical translation to pass through the point (1.5, 2). The smooth for each diagnostic plot was computed at 41 points spaced equally in [1,2]; i.e., 1.00, 1.025, . . . , 2.00. This interval covered 90% of the data. The simulations were summarized by the 41 pointwise means of the 1000 smooths for each correlation.

Figure 3 shows residual plots for the functional form for Z_2. The weighted smoothed GLM partial residual plot (long dash line) performs well; the mean smooth tracks the true functional form (solid line) closely, if not exactly, regardless of the covariate correlation. The smoothed martingale residual plot ignoring Z_2 (dotted line) performs well only with uncorrelated covariates. With highly correlated covariates, it shows substantial bias. The unweighted smoothed GLM partial residual plot (short dash line) has poor performance with both correlated and uncorrelated covariates. Simulation results for other models and smoothers lead to similar conclusions [37].

The second motivation for the weighted GLM partial residuals comes from the penalized likelihood approach of Hastie and Tibshirani [38]. They considered simultaneous estimation of the functional form of all the covariates. They assumed that the hazard is $\lambda_0(t)\exp\{\Sigma_{j=1}^{p} f_j(Z_j)\}$ where the f_j are unspecified smooth functions. The resulting partial likelihood cannot be maximized directly without leading to overfitting and identifiability pro-

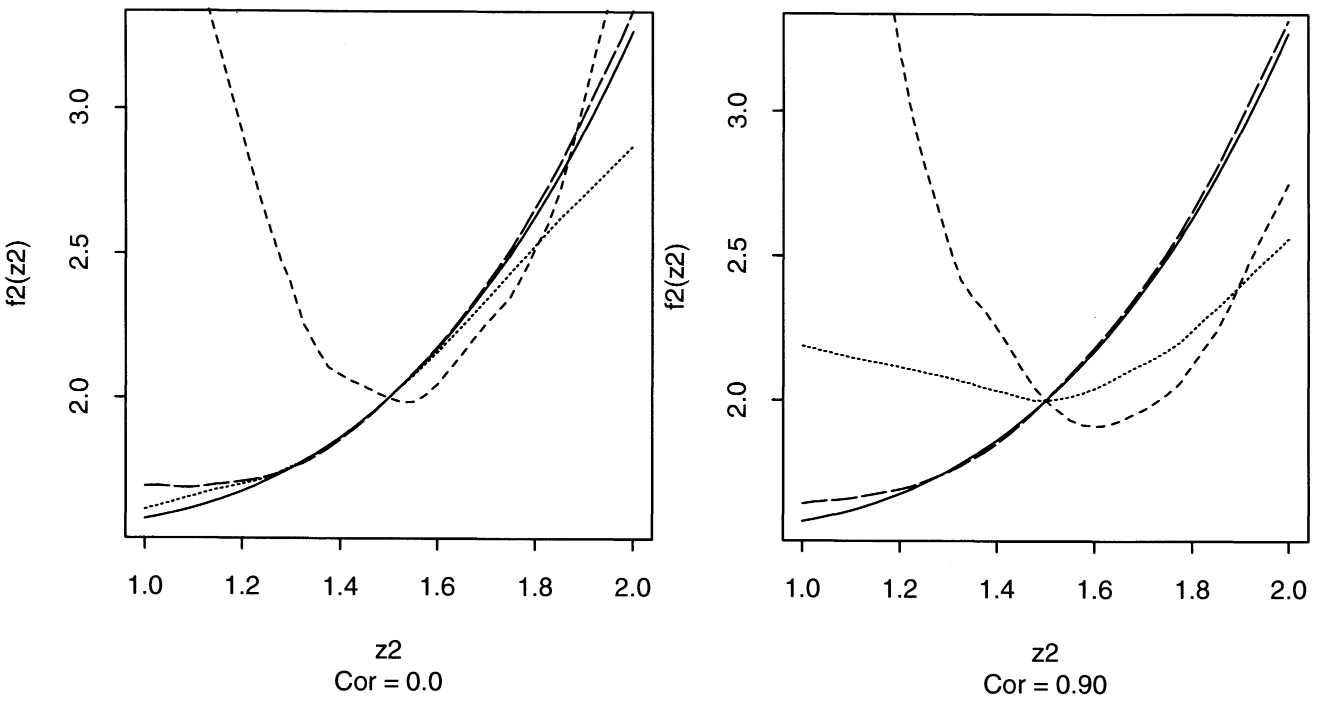


Figure 3. Summary of diagnostic techniques to estimate $f_2(Z_2)$ applied to simulated data from a model log-linear in Z_1 but not in Z_2. The long dashed line is the mean of 1000 smooths of weighted GLM partial residuals, the short dashed line is the mean smooths of unweighted GLM partial residuals, the dotted line is the mean of smooths of martingale residuals from a model without Z_2, and the solid line is $f_2(Z_2)$.

blems. Instead, they maximized the penalized partial likelihood with penalty function $\Sigma_{j=1}^{p}\lambda_j \int f_j''(s)^2 ds$, where $\lambda_i \geqslant 0$ $(i = 1, \ldots, p)$ are user-specified smoothing parameters. Let

$$\eta_i = \sum_{j=1}^{p} f_j(z_{ij}),$$

$$p_i(t) = \frac{Y_i(t)e^{\eta_i}}{\Sigma Y_j(t)e^{\eta_j}},$$

and

$$N.(t) = \sum_{i=1}^{n} N_i(t).$$

Then the derivatives of their partial likelihood are

$$\begin{aligned} \frac{\partial l}{\partial \eta_i} &= N_i(\tau) - \hat{\Lambda}(t_i)e^{\eta_i} \\ &= M_i. \\ -\frac{\partial^2 l}{\partial \eta_i^2} &= \int_0^{\tau} p_i(t)[1 - p_i(t)]dN_i(t) \\ &= E_i - \int_0^{\tau} p_i^2(t)dN.(t) \\ &= v_i. \end{aligned}$$

Their algorithm is a doubly iterated modified Newton–Raphson scheme involving repeated application of a weighted cubic smoothing spline to the scaled martingale residual, M_i/v_i, plus the current estimated functional form $\hat{f}_i$, with weights given by the v_i's. The first step is effectively the weighted smooth of the GLM partial residual plot proposed here, but with a particular smoother and with weights that differ from $\hat{E}_i$ by a factor of $O\left(\frac{1}{n}\right)$.

Iteration of our procedure should lead to approximately the same solution as Hastie and Tibshirani's method. An advantage of our approach is that special regression software is not necessary. Related techniques for estimating smooth f_i's include the local full likelihood approach of Gentleman and Crowley [39] and the penalized regression splines with a moderate number of knots from Gray [40]. These also require special software, since the optimization is done within the kernel of partial likelihood.

We apply our technique to a malignant melanoma data set [2], a historically prospective clinical study of 205 patients with malignant melanoma operated on at Odense University Hospital in 1962–1977 and followed until the end of 1977. The survival endpoint was death from malignant melanoma, and there were 57 such deaths. Andersen et al. [2] considered a Cox model with three predictors — two binary indicators, gender, and the presence or absence of ulceration in the tumor — and one quantitative predictor, tumor thickness (mm). Their analysis concluded with a model omitting gender, stratified on ulceration, and with tumor thickness as a proportional hazards covariate entered on the log scale. They decided on the log scale by starting with a model with tumor thickness untransformed. Adding other functions of tumor thickness, namely, $Z \times I\ (Z \geqslant 2\,\text{mm})$ and $Z \times I\ (Z \geqslant 5\,\text{mm})$, produced a statistically significant improvement in fit and suggested that the influence of thickness on the log hazard was a concave function. A model with log (thickness) was not significantly improved by adding other functions of thickness.

We used the weighted smooth GLM partial residuals to examine the functional form for tumor thickness in a model stratified on ulceration. We considered two models, one with thickness and the other with log (thickness). Figure 4 presents the residual plots for the two models. As is typical for censored failure time data, the partial residuals are heavily skewed with long right-hand tails. In fact, a few very large residuals were omitted from each plot. A weighted linear loess smooth with three equivalent degrees of freedom was applied. The functional form as estimated from each model (linear in thickness for the first model and linear in log thickness for the second) was also superimposed. The residual smooth for the linear model departs markedly and systematically from the linear fit, suggesting that a concave transformation would be better. The smooth for the log-linear model is closer to the linear fit. The partial residual plots lead to the same conclusion as Andersen et al. [2].

We also included the result of two additional iterations, where the smoothed weighted partial residual plus the fit from iteration i was used as the covariate for iteration $(i + 1)$. In this data set, as in several others examined by the authors, the iterated fit leads to the same conclusion as the initial smoothed residual plot.

Conclusion and summary

We have presented simple graphical techniques for assessing the two key assumptions of proportional hazards regression models for counting process data. Smoothed plots of suitably scaled Schoenfeld residuals can reveal departures from proportional hazards, and smoothed plots of martingale residuals can reveal departures from the postulated functional form. These techniques have two major advantages. The first is that they are computationally simple. They require only the Schoenfeld and martingale residuals, standard summary statistics from the Cox model such as the information matrix, and a plotting routine with scatterplot smoothing. The second advantage is that the plots do not simply show problems with the model but also suggest remedies. The smoothed scaled Schoenfeld residual plot suggests

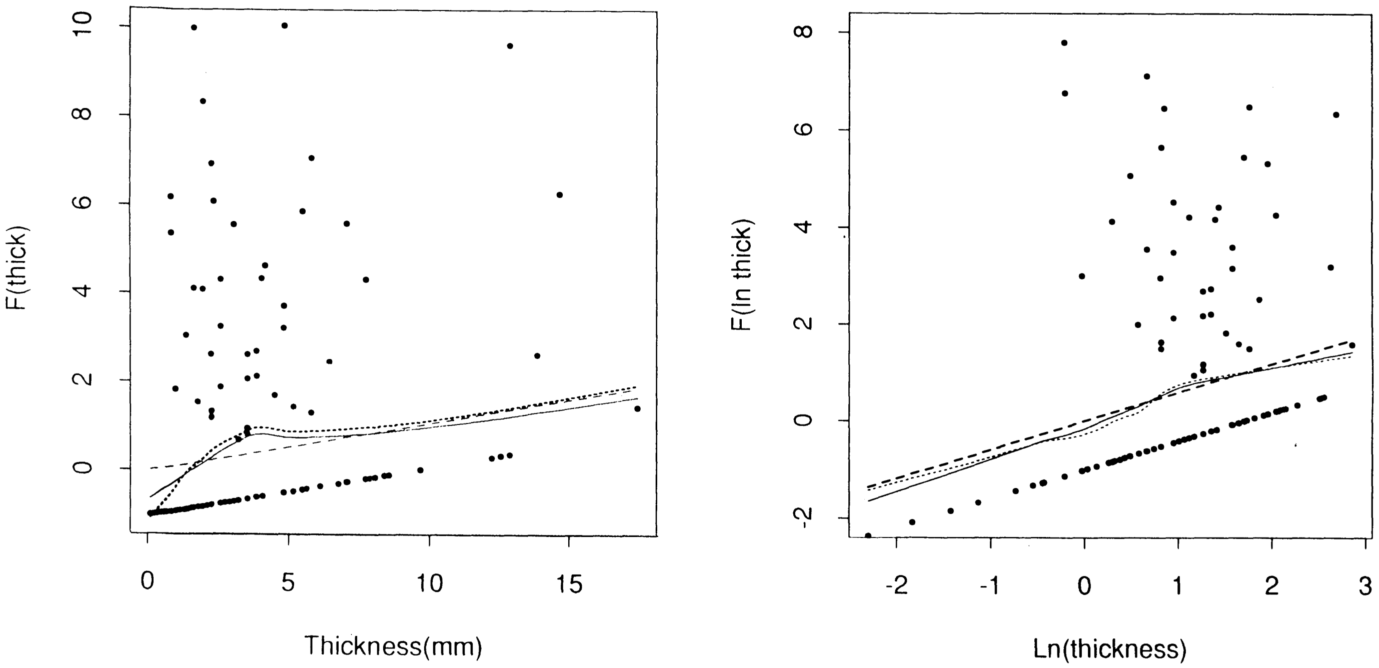


Figure 4. GLIM partial residual plots for the covariate tumor thickness in the malignant melanoma data. The left panel is from a model linear in tumor thickness, and the right panel is a model linear in log (thickness).

The solid line gives the weighted smoothed GLIM partial residual, the dashed line gives the estimate based on the model, and the dotted line gives the results of two additional iterations.

There were 17 residuals too large for the left panel and 18 too large for the right.

the functional form for a time-varying coefficient model. The smoothed scaled martingale residual plot suggests the functional form for a covariate in a proportional hazards model. Of course, the plots should be treated with due caution since, at base, they are based on one-step approximations to the true functions. At the cost of computational complexity, they could be iterated. Another drawback is that the two assumptions are treated in isolation. The martingale residual plot assumes that the proportional hazards assumption is met. Poorly specified functional form for covariates in a proportional hazards model may suggest lack of proportionality in diagnostic plots for proportionality. Further research into the relationships and interactions of these two plotting techniques will be beneficial. A possible approach involves simultaneously smoothing estimated hazard rates as functions of both time and covariates [41].

References

1. Fleming TR, Harrington DP (1991). *Counting Processes and Survival Analysis*. New York: J. Wiley Sons.
2. Andersen PK, Borgan O, Gill RD, Keiding N (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.
3. Kay R (1984) Goodness-of-fit methods for the proportional hazards model: a review. *Rev Epidemiol Santé Publ* 32:185–198.
4. Lin DY, Wei LJ (1991). Goodness-of-fit tests for the general Cox regression model. *Stat Sinica* 1:1–17.
5. Cox DR (1972). Regression models and life-tables (with discussion). *J R Stat Soc B* 34: 187–220.
6. Andersen PK, Gill RD (1982). Cox's regression model for counting processes: a large sample study. *Ann Stat* 10:1100–1120.
7. Breslow NE (1974). Covariance analysis of censored survival data. *Biometrics* 30:89–99.
8. Barlow WE, Prentice R (1988). Residuals for relative risk regression. *Biometrika* 75:65–74.
9. Schoenfeld D (1982). Partial residuals for the proportional hazards regression model. *Biometrika* 69:239–241.
10. O'Quigley J, Pessione F (1989). Score tests for homogeneity of regression effects in the proportional hazards model. *Biometrics* 45:135–144.
11. Schoenfeld D (1980). Chi-square goodness of fit tests for the proportional hazards model. *Biometrika* 67:145–153.
12. Moreau T, O'Quigley J, Mesbah M (1985). A global goodness-of-fit statistic for the proportional hazards model. *Appl Stat* 34:212–218.
13. Harrell F (1986). The PHGLM procedure. *SAS Supplemental Library User's Guide*, Version 5. Cary, NC: SAS Institute Inc.
14. Lin DY (1991). Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators. *J Am Stat Assoc* 86:725–728.
15. Nagelkerke NJD, Oosting J, Hart AAM (1984). A simple test for goodness of fit of Cox's proportional hazards model. *Biometrics* 40:483–486.
16. Jones MP, Crowley J (1989). A general class of nonparametric tests for survival analysis. *Biometrics* 45:157–170.
17. Jones MP, Crowley J (1990). Asymptotic properties of a general class of nonparametric tests for survival analysis. *Ann Stat* 18:1203–1220.
18. Harrington DP, Fleming TR (1982). A class of rank test procedures for censored survival data. *Biometrika* 69:533–546.

19. Tarone RE, Ware J (1977). On distribution-free test for equality of survival distributions. *Biometrika* 64:156–160.
20. O'Brien PC (1978). A nonparametric test for association with censored data. *Biometrics* 34:243–250.
21. Pettitt AN, Bin Daud I (1990). Investigating time dependence in Cox's proportional hazards model. *Appl Stat* 39:313–329.
22. Cleveland WS, Devlin SJ (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J Am Stat Assoc* 83:596–610.
23. Cleveland WS, Devlin SJ, Grosse E (1988). Regression by local fitting: mothods, properties, and computational algorithms. *J Econometr* 37:87–114.
24. Cleveland WS, Grosse E, Shyu WM (1992). Local regression models. In *Statistical Models in S*, JM Chambers, JJ Hastie (eds.). Pacific Grove, CA: Wadsworth and Brooks, 309–376.
25. Cleveland WS (1979). Robust locally-weighted regression and smoothing scatterplots. *J Am Stat Assoc* 74:829–836.
26. Hastie TJ, Tibshirani RJ (1990). *Generalized Additive Models*. London: Chapman and Hall.
27. Stablein DM, Carter WH Jr, Novak J (1981). Analysis of survival data with nonproportional hazard functions. *Controlled Clin Trials* 2:149–159.
28. Chang IS, Hsiung CA (1990). Finite sample optimality of maximum partial likelihood estimation in Cox's model for counting processes. *J Stat Planning Infer* 25:35–42.
29. Therneau T (1993). A package of survival functions for S, available from StatLib.
30. Becker RA, Chambers JM, Wilks AR (1988). *The New S Language*, Pacific Grove, Califormia: Wadsworth & Brooks/Cole.
31. Statistical Sciences, Inc. (1993). *S-Plus Training Manual: Introductory/Advanced, Version 3.1 for Unix*. Seattle, WA: Statistical Sciences, Inc.
32. Fleming TR, O'Fallon JR, O'Brien PC, Harrington DP (1980). Modified Kolmogorov–Smirnov test procedures with applications to arbitrarily censored data. *Biometrics* 36: 607–625.
33. Gill R, Schumacher M (1987). A simple test of the proportional hazards assumption. *Biometrika* 74:289–300.
34. Grambsch PM, Therneau TM (in press). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*.
35. Therneau TM, Grambsch PM, Fleming TR (1990). Martingale-based residuals for survival models. *Biometrika* 77:147–160.
36. McCullagh P, Nelder JA (1989). *Generalized Linear Models*, 2nd edition. New York: J. Wiley and Sons.
37. Grambsch PM, Therneau TM, Fleming TR (1994). Diagnostic plots to reveal functional form for covariates in multiplicative intensity models. Research Report 94–005, University of Minnesota Division of Biostatistics Research Report Series, Minneapolis, MN.
38. Hastie T, Tibshirani R (1993). Varying-coefficient models (with discussion). *J R Stat B* 55:757–796.
39. Gentleman R, Crowley J (1991). Local full likelihood estimation for the proportional hazards model. *Biometrics* 47:1283–1296.
40. Gray RJ (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J Am Stat Assoc* 87:942–951.
41. Gray RJ (1990). Some diagnostic methods for Cox regression models through hazard smoothing. *Biometrics* 46:93–102.

6. A review of tree-based prognostic models

Michael LeBlanc and John Crowley

Introduction

Tree-based methods are adaptive nonparametric statistical procedures that are gaining popularity in applied fields such as artificial intelligence, pattern recognition, and medicine. An important attraction of tree-based models is their interpretation in terms of a partition of the covariate space and their binary decision tree representation.

Tree-based regression models are typically constructed by partitioning the data recursively into groups that minimize some measure of impurity — for instance, the residual sum of squares for a continuous response variable or binomial deviance for a binary response variable. Alternatively, the data are split based on some measure of dissimilarity between two groups appropriate for the response distribution, such as a two-sample test statistic. The splitting of the data continues until there are only a few observations in each region, leading to a model that overfits the data. Then the tree is pruned back, and the 'best' model is selected.

Tree-based models were first introduced by Morgan and Sonquist [1]. However, they gained popularity due to the work of Breiman et al. [2], who developed the Classification and Regression Tree (CART) algorithm. The CART algorithm includes important improvements to the methodology, including first growing a large tree, an optimal pruning algorithm, and cross-validation to estimate prediction error and to select the tree size. Statistical software for the CART algorithm and other flexible tools for tree-based models in the S statistical language [3] by Clark and Pregibon [4] have further increased interest in the methodology.

Whereas the methodology was developed for categorical or continuous outcome data, there is also interest in using this methodology for censored survival data, where investigators frequently want to find groups of patients with differing prognosis. While Cox's [5] proportional hazards model is a flexible tool for the study of covariate associations with survival time, it does not directly lead to models for prognostic groups. It seems reasonable that a model for prognostic groups should describe groups of patients with relatively homogeneous survival probabilities. Such a model is piecewise constant over

Peter F. Thall (ed.), RECENT ADVANCES IN CLINICAL TRIAL DESIGN AND ANALYSIS.
Copyright © 1995 Kluwer Academic Publishers, Boston. All rights reserved. ISBN 978-1-4613-5830-5.

regions of the covariate space with a single survival function corresponding to each region. Some examples of tree-based methods for forming prognostic groups with survival data are given in Albain et al. [6], Ciampi et al. [7], and Kwak et al. [8].

In addition to identifying prognostic groups, tree-based methods have other desirable statistical attributes. These include invariance to monotonic transformations of the predictor variables, flexibility to adjust to nonlinear or nonadditive covariate effects, and the potential to include good methods to deal with missing covariate values.

Much of the recent methodological work on tree-based methods for survival data has concentrated in three main areas:

- Tree-based methods have been developed that use some measure of impurity suitable for censored survival data. This enables one to incorporate most of the 'good' engineering aspects of the CART algorithm [9–12]. Davis and Anderson [9] use an exponential model loglikelihood to define impurity for a node. LeBlanc and Crowley [11] use an approximation to the full likelihood for the proportional hazards model. Gordon and Olshen [10] use L_p and L_p Wasserstein metrics [13], which focus on vertical and horizontal differences between distribution functions, respectively. Therneau et al. [12] propose growing trees on residuals from the Cox [5] model.
- Methods that focus on the separation between nodes have also been developed. The logrank test statistic is used to partition the data recursively and to develop the models [14,15]. LeBlanc and Crowley [16] develop an optimal pruning algorithm based on between-node test statistics, and they propose various resampling techniques to aid in the selection of tree size.
- Graphical and analytical methods for studying a single split in a tree-based model for survival data have also been studied. Graphical methods have been discussed in LeBlanc [17] and LeBlanc and Crowley [18]. Asymptotic properties for splitting data based on the logrank test statistic have been studied by Jesperson [19], Lausen and Schumacher [20], and LeBlanc and Crowley [18].

In this chapter, we discuss some of the general methodological aspects of tree-based models for survival data as cited in the first two points above. Finally, we show an example based on data from a clinical trial for myeloma.

Data and model

The data are assumed to consist of failure times and covariates that may be associated with failure times. An observation is distributed as the vector (T,Δ,X), where T is the time under observation, Δ is an indicator of failure, and X is a vector of p covariates. The learning sample consists of the set of independent observations $\{(t_i,\delta_i,x_i): i = 1, 2, \ldots, N\}$.

The techniques for growing tree-based models are described in the following section. However, a tree-based model for survival can be expressed by a step-function regression model

$$S(t \mid x) = \sum_{h \in \tilde{T}} S_h(t) I\{x \in B_h\},$$

where $\tilde{T}$ is the set of terminal nodes (nodes with no daughter nodes) of a binary tree T and $\{B_h, h \in \tilde{T}\}$ forms a partition of the predictor space. The function $S_h(t)$ is the survival curve corresponding to region B_h. Therefore, the mean or median function corresponding to the survival model would be a simple piece-wise constant function over the covariate space. Note that one could choose a simple parametric model such as the exponential model $S_h(t) = \exp(-\lambda_h t)$ (e.g., [9]) or a more general proportional hazards model where $S_h(t) = S_0(t)^{\theta h}$ and $S_0(t)$ is a nonparametric baseline survival function (e.g., [11]). Other general models may be useful in some situations. However, models should allow computationally efficient parameter estimation because many partitions need to be evaluated to construct a tree-based model.

Growing a tree

A tree-based model is grown by first splitting the covariate space into two regions and the data into two groups. The same splitting rule is applied recursively to each of the resulting regions until a large tree has been grown.

Typically, splits on a single covariate are used because they are easier to evaluate and interpret. For an ordered covariate, splits are of the form '$X_j \leqslant c$' or '$X_j > c$' and for a nominal covariate splits are of the form '$X_j \in S$' or '$X_j \in \bar{S}$,' where $B = \{b_1, b_2, \ldots, b_r\}$ and S is a subset of B. All possible splits are evaluated for each of the covariates, and the covariate and split point resulting in the greatest reduction in impurity is chosen. In addition, there is typically a rule limiting the minimum number of observations in a node to control the amount of adaptiveness of the algorithm.

A useful measure of impurity for a node is the deviance corresponding to the data for the node and the assumed survival model. The deviance for node h is defined to be

$$R(h) = 2\{L_h(\text{saturated}) - L_h(\hat{\theta}_h)\},$$

where L_h(saturated) is the log-likelihood for the saturated model that uses one parameter for each observation, and $L_h(\hat{\theta}_h)$ is the maximized log-likelihood for node h with maximum likelhood estimate $\hat{\theta}_h$. For instance, the exponential model deviance for node h is

$$R(h) = \sum 2\left[\delta_i \log\left(\frac{\delta_i}{\hat{\lambda}_h t_i}\right) - (\delta_i - \hat{\lambda}_h t_i)\right],$$

where $\hat{\lambda}_h$ is the maximum likelihood estimate of the hazard rate in node h.

The improvement for split s at node h into left and right daughter nodes $l(h)$ and $r(h)$ is

$$G(s,h) = R(h) - [R(l(h) + R(r(h))].$$

Alternatively, the best split can be based on maximizing the separation in survival times between two groups. These methods use the logrank test statistic

$$G(s,h) = G_0(s,h)^2/V(s,h), \tag{1}$$

where

$$G_0(s,h) = \int_0^{\infty} w(u) \frac{Y_1(u)Y_2(u)}{Y_1(u) + Y_2(u)} (d\hat{\Lambda}_1(u) - d\hat{\Lambda}_2(u)),$$

and where $\hat{\Lambda}_j(t)$ and $Y_j(t)$ are the Nelson [21] cumulative hazard estimator and number of individuals at risk for each of the two groups defined by the split s at time t. $V(s,h)$ is an estimate of the variance of the logrank numerator, $G_0(s,h)$. In figure 1, an example of a single split on a continuous covariate is given in the left panel, and the right panel shows the Kaplan–Meier [22] estimates of the survival curves for two of the groups data corresponding to the regions generated by the split, where the lower curve corresponds to data with $x_1 < 1.64$ and the upper curve to data with $x_1 \geqslant 1.64$.

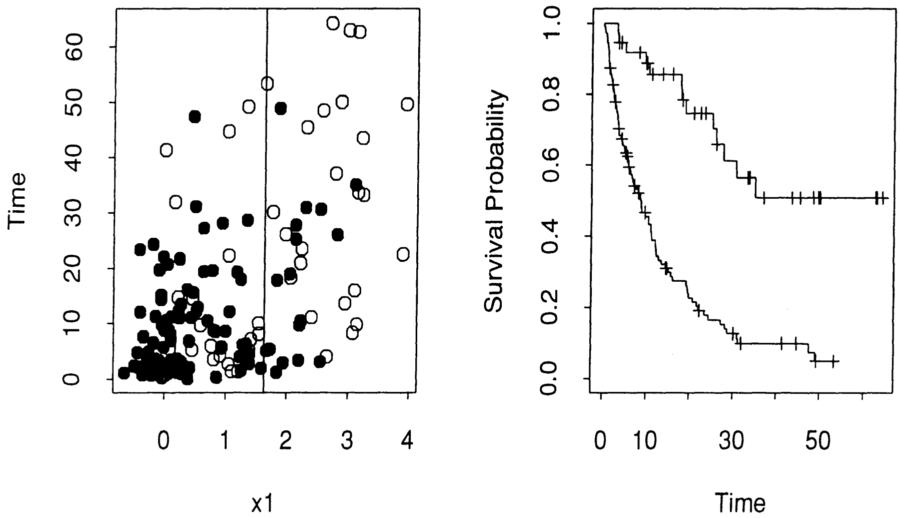


Figure 1. Hypothetical split on a covariate (left panel) and survival function estimates for each group (right panel). In the left panel, open circles represent censored times and shaded symbols represent deaths. In the right panel, '+' symbols represent censored times.

It is also important that the splitting statistic can be efficiently updated for all possible split points for continuous covariates. Simple updating algorithms exist for exponential deviance and approximations to the logrank test statistic [9,16].

The binary splitting continues until a relatively large tree has been grown. Figure 2 shows a partition and tree representation for a small tree-based model developed on a data set with two covariates.

Pruning and tree selection

Several different methods have been proposed for pruning trees and for the selection of one or a few models. The methods that use within-node error or deviance usually adopt the CART pruning algorithm directly.

Within-node methods

In the CART algorithm, the cost-complexity measure

$$R_\alpha(T) = \sum_{h\in\tilde{T}} R(h) + \alpha|\tilde{T}|$$

is used to assess the performance of a tree-based model, where $\tilde{T}$ is the set of terminal nodes in a binary tree T, $|\tilde{T}|$ is the number of terminal nodes, α

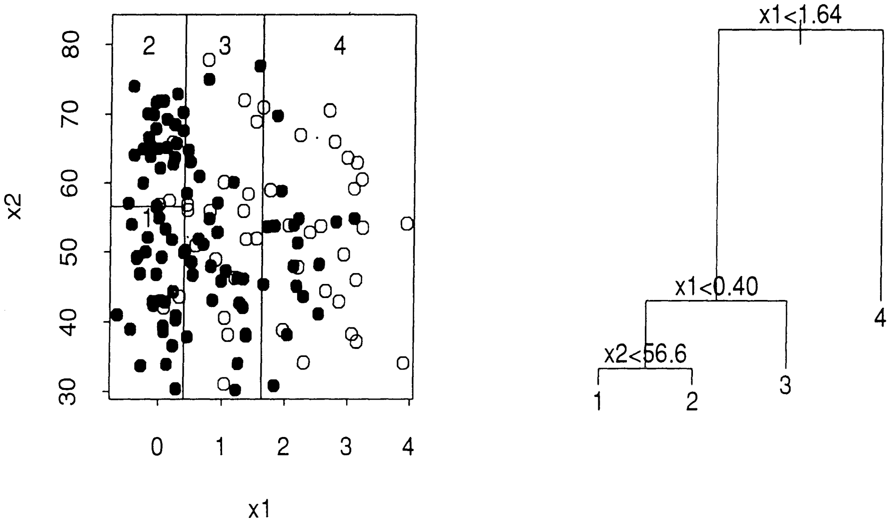


Figure 2. Partition of a two-dimensional covariate space (left panel) and binary tree representation (right panel). The numbers at the bottom of the tree correspond to labels for the regions defined.

is a nonnegative parameter, and $R(h)$ is the impurity or estimated cost of node h defined above.

A subtree (a tree obtained by removing branches) T_1 is an optimally pruned subtree for any penalty α of the tree initially grown if

$$R_\alpha(T_1) = \min_{T' \leqslant T} R_\alpha(T'),$$

where '$\leqslant$' means 'is a subtree of,' and T_1 is the smallest optimally pruned subtree if $T_1 \leqslant T''$ for every optimally pruned subtree, T''.

The cost-complexity pruning algorithm efficiently obtains the optimally pruned subtree for any α. This algorithm finds the sequence of optimally pruned subtrees by repeatedly removing branches for which the average reduction in impurity per split in the branch is small. The process yields a nested sequence of subtrees. The cost-complexity pruning algorithm is necessary for finding optimal subtrees because the number of possible subtrees grows rapidly as tree size increases. A tree with only 32 terminal nodes can have almost 460,000 potential subtrees [2].

After the pruning algorithm yields a sequence of trees; the selection of the best tree is guided by a cross-validation estimate of deviance. The data, $\mathscr{L}$, are divided up into V test samples $\mathscr{L}_v$ and training samples $\mathscr{L}_{(v)} = \mathscr{L} - \mathscr{L}_v$, $v = 1, \ldots, V$ of about equal size. Trees are grown with each of the training samples $\mathscr{L}_{(v)}$; each test sample $\mathscr{L}_v$ is sent down the tree to estimate the deviance using the parameter estimates from the training sample $\mathscr{L}_{(v)}$. The results are summed over the V test samples to obtain the cross-validation estimate of deviance. The tree that minimizes the cross-validation estimate of deviance (or a slightly smaller tree) is selected. The CART algorithm includes the '1 standard error rule,' which selects the smallest tree that does not perform significantly worse than the tree that minimizes the cross-validated estimate of prediction error. In addition to giving a smaller tree that is easier to interpret, the trees selected by the '1 standard error rule' have less variability in size than those selected without the rule.

Between-node methods

Methods that only use between-node separation [14,15] adopt more ad hoc methods for pruning trees and selecting tree size than the cost-complexity pruning and cross-validation used in the CART algorithm. However, for trees that are based on between-node separation, LeBlanc and Crowley [16] develop an optimal pruning algorithm similar to the cost-complexity pruning algorithm of CART. 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 (or splits) T.

One can interpret $G(T)$ as the amount of prognostic structure represented by the tree-based model. Such an interpretation can be motivated by considering the logrank test statistic in equation (1) as a standardized distance between empirical hazard functions of adjacent nodes in the tree.

A tree T_1 is an optimally pruned subtree of T for complexity parameter α if

$$G_\alpha(T_1) = \max_{T' \leqslant T} G_\alpha(T'),$$

and it is the smallest optimally pruned subtree if $T_1 \leqslant T''$ for every optimally pruned subtree, T''.

A pruning algorithm analogous to the cost-complexity pruning algorithm leads to the best tree for any α, just as with the cost-complexity pruning algorithm. The algorithm repeatedly prunes off branches with smallest average logrank test statistics in the branch. Either bootstrap or permutation sampling methods are used to select the tree size. Some permutation sampling methods are presented in the following example.

Example: prognostic groups for myeloma

This example is based on data from 614 patients who were entered on a randomized clinical trial of the Southwest Oncology Group between 1982 and 1987. The subset of the patient characteristics and laboratory values considered here are *sex*, sex; *age*, age; *pfs*, performance status; *cal*, serum calcium; *cre*, serum creatinine; *alb*, albumin; *sb2*, serum β_2 microglobulin; *pro*, percent proplasmacytes; *plb*, percent plasmablasts; *pho*, acid phosphate; *glu*, beta glucuronidase; *mpc*, percent mature plasma cells; *hgs*, hemoglobin. Analyses of other subsets of variables and earlier versions of the data are presented in Saeed et al. [23] and LeBlanc and Crowley [11].

A tree-based model was grown on the data using an algorithm based on between-node separation with logrank test statistics. The full unpruned tree has 20 terminal nodes and is not displayed. A nested sequence of optimally pruned subtrees was generated by the pruning algorithm referenced in the previous section and developed in LeBlanc and Crowley [16].

Permutation samples are used to determine approximate p-values and degrees of freedom for an adaptively chosen split. We permute the responses $\{(T_i,\delta_i)\colon i \in B_h\}$ over the covariate values within the node and calculate the maximal text statistic, G_b, for each permutation sample $b = 1, \ldots, B$. Under the assumption that the censoring and survival distribution are not associated with the covariates in the node, an approximate p-value for an adaptively chosen split is

$$P_h = \frac{1}{B+1}\left[\sum_{k=1}^{B} I\{G_b(h) \geqslant G(h)\} + 1\right],$$

where $G(h)$ is the maximal test statistic for the original sample.

We define approximate degrees of freedom for an adaptively chosen split as

$$d_h = 2 \times \frac{1}{B}\sum_{b=1}^{B} G_b(h).$$

Here, d_h is an estimate of the expected value of the logrank test statistic if the survival and censoring times are not associated with the covariates, for observations in node h. Note that $d_h \approx 1$ if there were no adaptive selection of the split point.

A model performance analogous to the Akaike Information Criterion (AIC) [24], defined as

$$GC(T) = G(T) - 2 \times \sum_{h \in T} d_h,$$

is used to assess the performance of the tree.

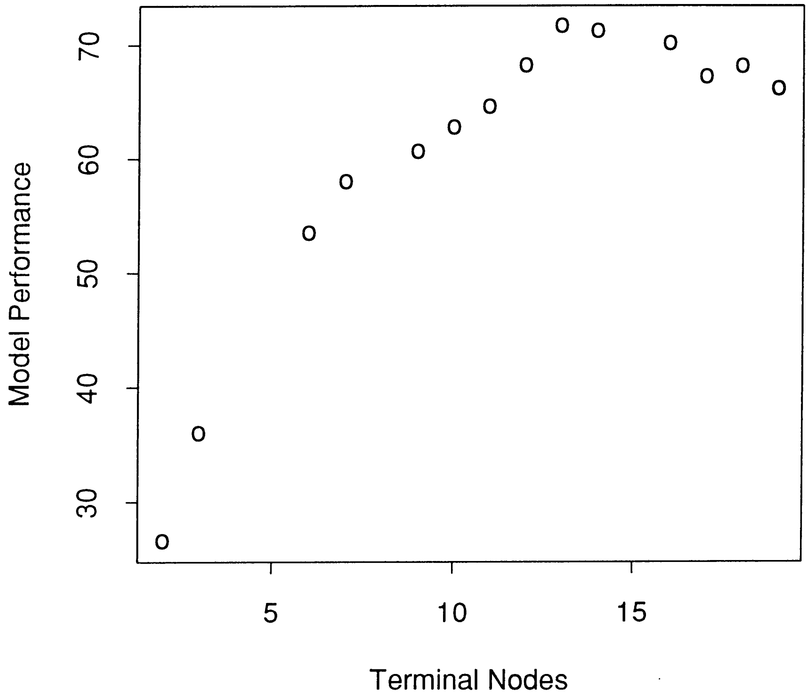


Figure 3. Model performance $GC(T)$ and the pruned sequence of trees for the myeloma data. The tree with 13 terminal nodes yields the best model performance.

A plot of $GC(T)$ for the nested sequence of optimally pruned subtrees is given in figure 3. The tree with best model performance has 13 terminal nodes and is presented in figure 4. In addition to the logrank test statistic, degrees of freedom and approximate p-values based on 1000 permutation samples are presented below each split. The logarithm of the relative risk estimates and the number of observations are presented below each terminal node.

The development of a small number of prognostic groups was an important goal of the analysis. Hence, the relative risk estimates given below each terminal node shown in figure 4 were used to order the nodes from best to worst prognosis:

3,10,5,12,1,4,6,9,8,2,11,13,7

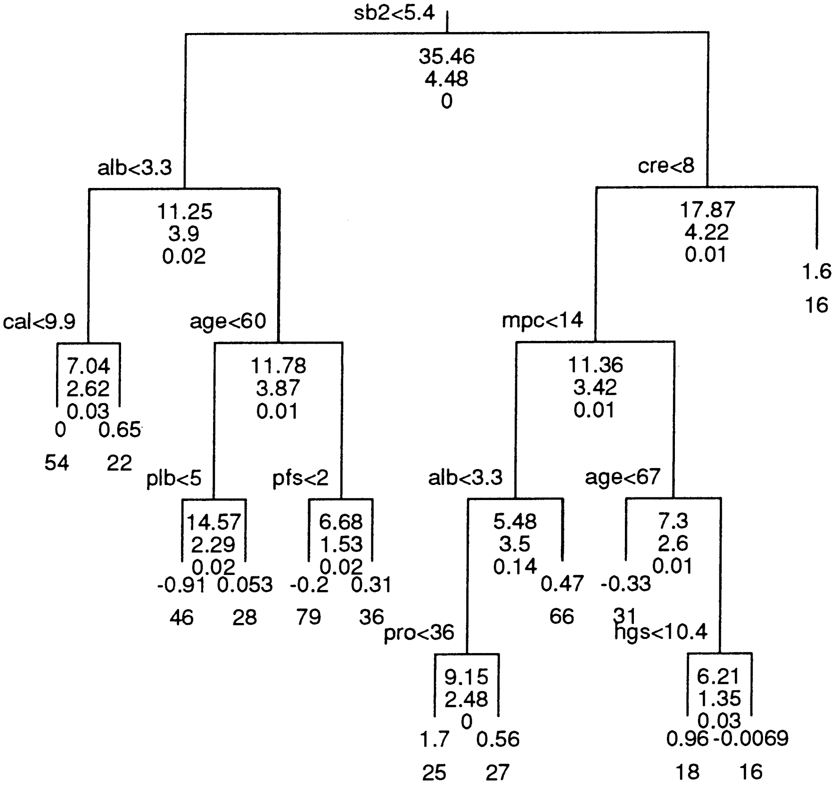


Figure 4. Pruned tree for the myeloma data. The split value, the logrank test statistic, approximate degrees of freedom, and approximate p-value (based on 1000 permutation samples) are given for each split. The logarithm of relative risk estimates (with the leftmost node as baseline) and the number of observations are given below each terminal node. The number of observations at the terminal nodes is less than the total sample size because of missing values for some covariates.

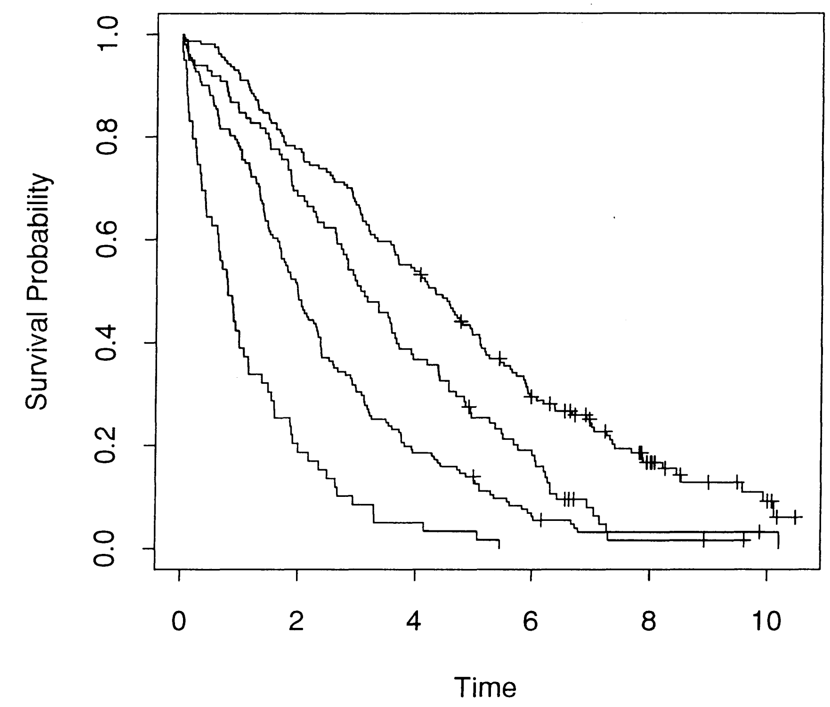


Figure 5. Survival function estimates corresponding to four-groups stratification of the pruned tree. The nodes for the groups for best to worst prognosis are (3,10,5), (12,1,4), (6,9,8,2), and (11,13,7), where the numbers indicate the node indices (numbered from left to right in figure 4).

(nodes are numbered consecutively from left to right in figure 4). These nodes were divided to construct four groups of approximately equal sample size. The survival function estimates and the nodes labels are given in figure 5 and its legend.

For the data presented, it was also of interest to determine close-competitor splits for the first split of the data. A simple graphical tool gives logrank test statistics for all possible splits on four of the covariates, namely, serum β_2 microglobulin, serum creatinine, percent mature plasma cells, and age. The results are presented in figure 6. The best competitor split on a different covariate is *age* $\leq$ 67. The same technique could be used to assess the 'stability' of splits and to look for competitor splits for other parts of the tree.

Software for growing the tree-based models presented in this section and other interactive graphical tools for developing partitions for survival have been written in the S statistical language and are available from the first author.

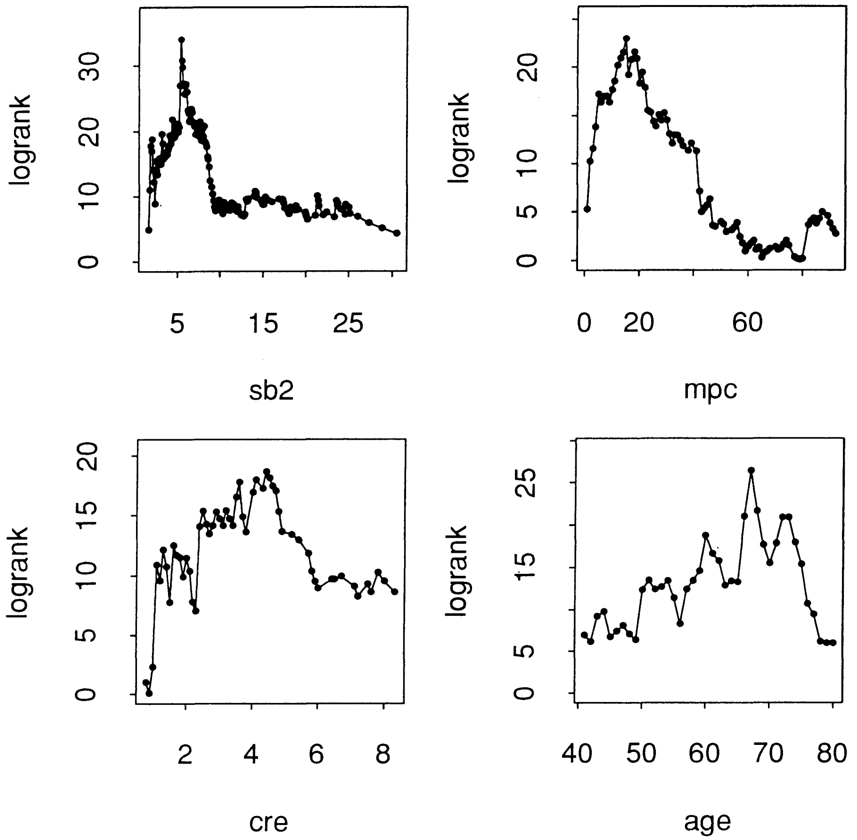


Figure 6. Logrank test statistics for competitor first splits of the data on four variables: serum β_2 microglobulin, percent mature plasma cells, serum creatinine, and age. Upper right panel shows the optimal split at *sb2* < 5.4.

Acknowledgments

The multiple myeloma data are used with the permission of Dr. Sydney Salmon, Chair of the Myeloma Committee of the Southwest Oncology Group. The first author was supported by the Natural Sciences and Engineering Research Council of Canada and the second by the US NIH through S U10 CA 38926 and S P01 CA 53996.

References

1. Morgan J, Sonquist J (1963). Problems in the analysis of survey data and a proposal. *J Am Stat Assoc* 58:415–434.

2. Breiman L, Friedman J, Olshen R, Stone C (1984). *Classification and Regression Trees*. Wadsworth International Group.
3. Becker R, Chambers J, Wilks A (1988). *The New S Language*. Wadsworth International Group.
4. Clark L, Pregibon D (1992). Tree-based models. *In Statistical Models in S*. Wadsworth International Group.
5. Cox DR (1972). Regression models and life tables. *J R Stat Soc B* 34:187–200.
6. Albain K, Crowley J, LeBlanc M, Livingston R (1990). Determinants of improved outcome in small cell lung cancer: an analysis of the 2580 patient Southwest Oncology Group data base. *J Clin Oncol* 8:1563–1574.
7. Ciampi A, Thiffault J, Nakache J-P, Asselain B (1986). Stratification by stepwise regression, correspondence analysis and recursive partition. *Comput Stat Data Anal* 4:185–204.
8. Kwak LW, Halpern J, Olshen RA, Horning SJ (1990). Prognostic significance of actual dose intensity in diffuse large-cell lymphoma: results of a tree-structured survival analysis. *J Clin Oncol* 8:963–977.
9. Davis R, Anderson J (1989). Exponential survival trees. *Stat Med* 8:947–962.
10. Gordon L, Olshen R (1985). Tree-structured survival analysis. *Cancer Treat Rep* 69: 1065–1069.
11. LeBlanc M, Crowley J (1992). Relative risk regression trees for censored survival data. *Biometrics* 48:411–427.
12. Therneau T, Grambsch P, Fleming T (1990). Martingale based residuals for survival models. *Biometrika* 77:147–160.
13. Shorack G, Wellner J (1986). *Empirical Processes and Applications to Statistics*. New York: John Wiley and Sons.
14. Ciampi A, Hogg S, McKinney S, Thiffault J (1988). RECPAM: a computer program for recursive partition and amalgamation for censored survival data. *Comput Methods Programs Biomed* 26:239–256.
15. Segal M (1988). Regression trees for censored data. *Biometrics* 44:35–48.
16. LeBlanc M, Crowley J (1993). Survival trees by goodness of split. *J Am Stat Assoc* 88:457–467.
17. LeBlanc M (1990). Tree-based tools for survival data. *Proceedings of the XV International Biometrics Conference*, 123–133.
18. LeBlanc M, Crowley J (1995). Step function covariate effects in the proportional hazards model. To appear in *Canadian Journal of Statistics*.
19. Jesperson (1986). Technical report, University of Copenhager.
20. Lausen B, Schmucher M (1992). Maximally selected rank statistics. *Biometrics* 48:73–85.
21. Nelson (1969). Hazard plotting for incomplete failure data. *J. Qual. Technology* 1:27–52.
22. Kaplan E, Meier P (1958). Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 53:457–481.
23. Saeed SM, Stock-Novack D, Pohlod R, Crowley J, Salman SE (1993). Prognostic correlation of plasma cell acid phosphatase and beta-glucoronidase in multiple myeloma. *Blood* 81:869–870.
24. Akaike H (1974). A new look at model identification. *IEEE Trans Automat Control* 19:716–723.

7. Decision analysis and Bayesian methods in clinical trials

Donald A. Berry

Introduction

The standard statistical approach to designing and analyzing clinical trials is frequentist. A purpose of this chapter is to describe a Bayesian approach as an alternative, or perhaps as a supplement. The two approaches have focuses so different that they can be viewed as distinct disciplines. And yet both deal with empirical evidence and both use probability, so the distinction is poorly understood by nonstatisticians. The distinction is further blurred when frequentists and Bayesians act and think alike — which they are wont to do, despite 'anti' rhetoric coming from both sides. Both approaches have good characteristics. The most important advantage of the Bayesian approach is attitude, which is consistent with the scientific method.

An example will help show an important difference in the approaches. A clinical trial has been carried out to compare an experimental with a standard therapy. One type of frequentist conclusion is a p-value: the probability of results more extreme than those observed in the trial given the null hypothesis H_0 that the two therapies have identical effectiveness. In symbols, $p\text{-value} = P(\text{DATA} \mid H_0)$, where DATA refers to the totality of data more extreme than that observed. (The reason for including 'more extreme' data is that the probability of any particular observation is usually tiny, regardless of the hypothesis assumed.) Another type of frequentist conclusion is the probability of results more extreme than those observed in the trial under the particular alternative hypothesis that the new therapy indeed has a particular benefit — 20% say. These probabilities depend on the results of the trial at hand and so are descriptors of the trial results. (Which data are more extreme depends on the way the trial was conducted — on the stopping rule, for example [1,2] — but this matter will not be addressed in this chapter.)

A Bayesian conclusion is the probability that the experimental and standard therapies are equally effective given the results of the trial at hand — in symbols, $P(H_0 \mid \text{data})$. Another Bayesian conclusion is the probability that the experimental therapy is at least 20% more effective than the

Peter F. Thall (ed.), RECENT ADVANCES IN CLINICAL TRIAL DESIGN AND ANALYSIS.
Copyright © 1995 Kluwer Academic Publishers, Boston. All rights reserved. ISBN 978-1-4613-5830-5.

standard, again given the trial's results. These are called *posterior probabilities* in that they apply *after* the trial. Probabilities that refer to future observations on individual patients or on finite sets of patients (again given the trial's results) are called *predictive probabilities*. So Bayesians use probabilities of hypotheses given the results while frequentists use probabilities of sets of results given hypotheses.

The results of the trial are known, and so it seems reasonable to condition on them. The Bayesian approach is so natural that users interpret frequentist measures as though they were Bayesian probabilities. For example, if the p-value for a therapy comparison is small, people tend to conclude that there probably is a difference in therapies. It is difficult to give a p-value any other interpretation!

Bayes' theorem

Calculating a posterior probability of any hypothesis H requires Bayes' theorem:

$$P(H \mid \text{data}) = P(\text{data} \mid H)/P(\text{data}).$$

In terms of a parameter — say w, which we can think of as the difference in effect between experimental and standard therapy — Bayes' theorem says that the posterior probability or density of any w is

$$p(w \mid \text{data}) = P(\text{data} \mid w)p(w)/P(\text{data}).$$

The factor $P(\text{data} \mid H)$ or $P(\text{data} \mid w)$ is the *likelihood function* evaluated at H or w. So Bayes' theorem relates the conditional density $p(w \mid \text{data})$ of a parameter w with its unconditional density $p(w)$. Since the latter depends on information present *before* the experiment, it is a *prior* probability. Think of $1/P(\text{data})$ as a factor that makes the total probability equal to 1 when adding over all possible w's — that is, the denominator $P(\text{data})$ is the sum (or integral) of the numerator over all w's. So rewriting Bayes' theorem:

$$\text{posterior.probability} \propto \text{likelihood} \times \text{prior.probability},$$

where $\propto$ means 'proportional to.' Bayes' theorem provides a formalism for learning: That's what I thought before (prior), this is what I just saw (likelihood), so here's what I now think (posterior) — and I may change my views tomorrow.

In a Bayesian analysis, prior information about a parameter w is assessed as a probability distribution on w. This distribution depends on the assessor, and so is *subjective*. Since posterior probabilities depend on prior probabilities, they too are subjective. A subjective probability can be calculated any time a person has an opinion. Counting ignorance as an opinion, though obviously a very weak one, this includes every setting.

The prior distribution of a parameter w includes all information that was available before the current trial. Interpreting this information in the context of the current trial and assessing prior probabilities is subjective. This is so even if some of the prior information itself arose from clinical trials. The patients in earlier trials may not have been similar to those in the current trial, and in any case we cannot know whether they were similar. So prior information has to be partially discounted (subjectively) when applied to the current setting.

Bayesian vs. frequentist approaches

Subjectivity is the principal objection to a Bayesian approach. Different people can draw different conclusions from the same results. While it is conforting when two analysts give the same answer, I regard subjectivity to be an advantage: differences of opinion are the norm in medicine and science generally, and so an approach that explicitly recognizes such differences is realistic.

Bayesian and frequentist probabilities are inverses of each other in the sense that the roles of the arguments and the conditions are exchanged. While this is an important distinction, it is not the most important. A more important difference is that a Bayesian conclusion depends on all available evidence, while the frequentist conclusion is restricted to the trial at hand. Bayesians assess evidence other than that in the trial and incorporate it into their analysis through a subjective assessment of prior probabilities. This makes the posterior probabilities relevant scientifically and medically, but it makes the Bayesian approach more difficult to use because it places an extra burden on an investigator and on consumers of experimental information.

The principal advantage of the Bayesian approach is its attitude toward evidence. Bayesian analysis is not data analysis per se, since its conclusions are not restricted to any particular data set. The Bayesian approach attempts to bring possibly different types of evidence to bear on questions of importance — questions such as whether a therapy is beneficial.

Decisions and utilities

Another important difference is that the Bayesian approach is decision oriented. A trial's results affect one's state of knowledge concerning the various therapies. But knowledge without application is specious. How should Ms. Smith be treated? Should another trial be designed? What type of trial? The answers depend on one's knowledge, but they also depend on the consequences of the various decisions. In a Bayesian approach, consequences are evaluated explicitly by associating a *utility* with each. Also, each consequence of a decision has a predictive probability. So the *utility of a decision* can be found by averaging utilities of consequences with respect to these predictive probabilities (see the section below on decision problems).

Much of the Bayesian clinical trial literature does not consider utilities or exploit the decision-analytic aspects of the Bayesian approach (see Spiegelhalter, Freedman, and Parmar [3] and many of their references). I will distinguish what I call a *fully Bayesian* approach from partial Bayesian approaches, without meaning to imply that less than fully Bayesian is less than good. A fully Bayesian approach is decision theoretic, and posterior probabilities are based on all available evidence, including evidence separate from the trial at hand. There are at least two ways to be less than fully Bayesian. First, one can calculate posterior distributions as data summaries without incorporating them into a decision analysis. Second, one can calculate posterior distributions using canonical prior distributions rather than prior distributions based on the available evidence. Bayesian approaches that are missing both of these characteristics are similar to the standard frequentist approach that focuses on data summary. But there are differences. The main one is flexibility: accumulating data from a clinical trial can be used to update Bayesian measures, independent of the design of the trial. Frequentist measures are tied to the design, and interim analyses must be planned for frequentist measures to have meaning. Its flexibility makes the Bayesian approach ideal for analyzing data from clinical trials.

The purpose of this chapter is to give examples of the use of Bayesian and decision theoretic methods. Other examples and further descriptions are given in Berry [4], Spiegelhalter, Freedman, and Parmar [3], Berry and Stangl [5], and their references.

A basic part of all Bayesian problems is the prior probability distribution of any unknowns. The next section gives a method for assessing prior probabilities.

Assessing probabilities

Subjective probability is based on degrees of belief. Consider an event whose occurrence is uncertain. How strongly an individual feels that this event will occur (or has occurred) depends on the individual as well as on the event. The purpose of this section is to show how to assess probabilities as degrees of belief. Anyone can assess his or her probabilities. For the purposes of designing clinical trials, the most important subjects are the investigators or other experts whose beliefs can be elicited. When results of clinical trials are published, the reader is the appropriate subject.

Calibrating for probability assessment

A basic requirement for assessing one's probabilities is the existence of a calibration scale. One must be able to imagine experiments in which the outcomes are exchangeable, in the following sense. Suppose the person gets to choose any one from among a set of outcomes and will receive $100 if the outcome chosen occurs. Outcomes in the set are *exchangeable* if the person is indifferent among the various outcomes, and in particular, would strictly prefer any one outcome over all others if the reward on it were increased by an arbitrarily small amount, say one cent. (Statements about preferences are always somewhat delicate because other people set the ground rules. The assessor should be able to imagine that there's no chicanery afoot.)

An experiment is a *calibration experiment* for someone if all outcomes of the experiment are exchangeable for that person. There are many candidates. But whether a particular experiment serves to calibrate depends on the assessor. What's required is that the assessor be indifferent and not that the probabilities are equal in a catholic sense. One convenient set of possibilities is to select a chip from a bowl that contains chips of the same size and shape.

Consider a specific setting. I'd like to know your probability that for a particular population, the average drop in diastolic blood pressure on a certain drug is less than 10 mmHg — call this event A. Consider a (calibration) bowl with one green and one red chip. I offer you the choice of getting $100 if a chip selected from the bowl is green or $100 if A is true; if you choose to select from the bowl and the chip is red or if you choose A and it turns out that A is false, then you receive nothing. Suppose you choose A; then I take this to mean that your $P(\mathrm{A})$ is at least 1/2. Now consider a (calibration) bowl with three green chips and one red chip. Again you get to choose between $100 if a chip selected from the bowl is green and $100 if A is true. If you now prefer the chip, then, taken together, your two answers mean that $1/2 \leqslant P(\mathrm{A}) \leqslant 3/4$. Proceeding in this way, each time halving the interval by doubling the total number of chips in the bowl, will give $P(\mathrm{A})$ sufficiently accurately.

There are several problems with this approach. The most obvious is that I cannot be serious about the money. And the problem is not that I have limited financial resources. I could be serious only if we could find out whether A is true. Luckily, in most practical circumstances, the assessor is motivated to take the assessment procedure seriously by imagining rewards in the comparison of events with calibration experiments.

Another problem is that the assessor soon faces very difficult decisions. By the time the bowl contains 16 chips, most assessors will have a hard time deciding between green and A. Again, luckily, a high degree of accuracy in specifying $P(\mathrm{A})$ is seldom required.

Probability distributions

Most problems require a probability distribution and not a single probability. The above process can be carried out for various events A of the form: average drop in diastolic blood pressure on the drug is less than x mmHg.

The prior distribution of drop in blood pressure can be determined within a specified accuracy by varying x. For example, figure 1 shows the result of carrying out this process for seven different changes in blood pressure. The assessor in question has probabilities 0 to 1/16, 1/16 to 2/16, 3/16 to 4/16, 5/16 to 6/16, 10/16 to 11/16, 14/16 to 15/16, and 15/16 to 1, for $X = -5, 0, 5, \ldots, 25$, respectively. The corresponding density is shown in figure 2, and a smoothed version is shown in figure 3. The advantage of the version in figure 3 over that in figure 2 is only cosmetic; in particular, it has no computational advantage.

Extreme values are of special importance in assessing probabilities. For example, suppose the drug is given to 20 subjects and the diastolic blood pressure of all 20 subjects *increases* by more than 10 mmHg (that is they drop by less than −10). Because the prior probability for the density in figure 2 gives all its probability to the right of −10, the posterior density

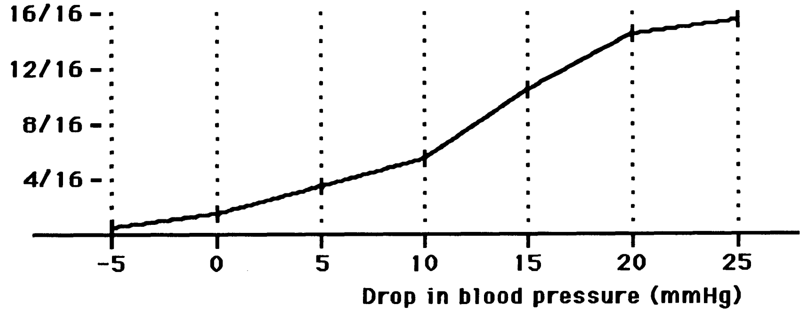


Figure 1. Subjective distribution function of change in blood pressure. Vertical lines show range indicated by the assessor. The curve drawn goes through the middle of the range at each point and so approximates the assessor's opinion.

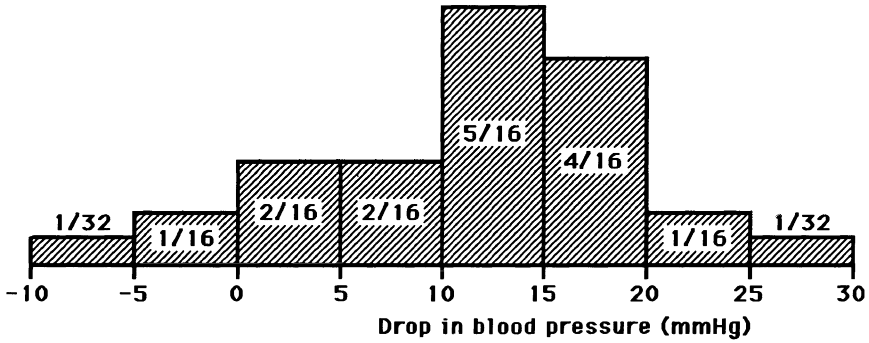


Figure 2. Subjective density function of drop in blood pressure estimated from figure 1.

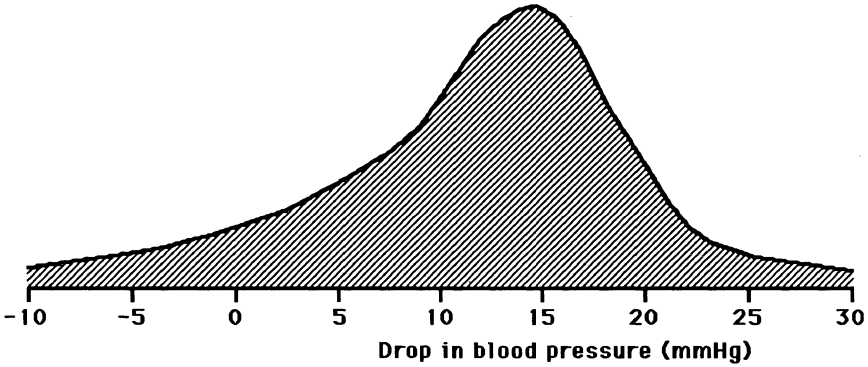


Figure 3. Smoothed version of density function from figure 2.

would give all its probability to the region just to the right of −10. This may not be appropriate and is likely due to a somewhat cavalier assessment. As a rule of thumb, be openminded to the extent that you give some probability with all possibilities, even if it is small.

Beta densities for proportions

A common type of problem is when there are two possible observations, such as success/failure or response/nonresponse or lives/dies. The parameter of interest is the proportion w of successes in the population. Suppose in a sample of n patients there are s successes and $f = n - s$ failures. Assuming independence of the observations conditional on w, the likelihood of w is proportional to

$$w^s(1 - w)^f.$$

Bayes' theorem says to multiply the prior density by this likelihood. There is a very convenient updating formula when the prior density is in the beta family. The beta(a,b) density is proportional to

$$w^{a-1}(1 - w)^{b-1}.$$

By Bayes' theorem, the posterior density is the product of these, namely,

$$w^s(1 - w)^f\, w^{a-1}(1 - w)^{b-1} = w^{a+s-1}(1 - w)^{b+f-1},$$

and is itself a beta density, with a replaced by $a+s$ and b replaced by $b+f$. Beta prior distributions for w are said to be *conjugate* because the new distribution of w stays in the beta family.

Assessing a prior density in the beta(a,b) family means finding a and b. Two assessments are required. The first is the assessor's probability of success on the first trial. This is the mean of the beta density — call it r — which has the simple form

$$r = \frac{\mathrm{a}}{a + b}.$$

For the second requirement, imagine that the first trial is a success, and assess the probability of success on the second trial:

$$r^{+} = \frac{a + 1}{a + b + 1}.$$

Solving simultaneously gives

$$a = \frac{r(1 - r^{+})}{r^{+} - r}, \quad \text{and} \quad b = \frac{(1 - r)(1 - r^{+})}{r^{+} - r}.$$

Consider an example. The most important prognostic factor in early breast cancer is the number of axillary lymph nodes testing positive during pathological review. The number of lymph nodes dissected during surgery varies. (This number and the nodes dissected may depend on clinical characteristics, but it is assumed here that the sampling is random and therefore that nodes selected are no different from those not selected.) The probability that any particular node is positive is 3%, and so $r = 0.03$. However, if the first one sampled tests positive, then the probability that the next is also positive increases dramatically to 20%: $r^{+} = 0.20$. Therefore,

$$a = \frac{0.03(1 - 0.2)}{0.2 - 0.03} = 0.14, \quad \text{and} \quad b = \frac{(1 - 0.03)(1 - 0.2)}{0.2 - 0.03} = 4.56.$$

So the prior density of w is beta(0.14,4.56), shown in figure 4.

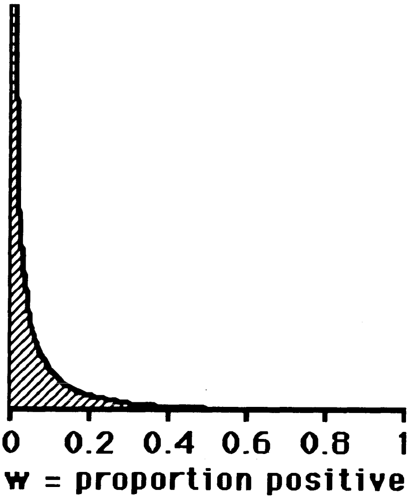


Figure 4. Beta(0.14,4.56) density for proportion of positive axillary lymph nodes. The mean is 0.03, which is also the probability of a positive lymph node when not conditioned on *w*.

Checking for consistency

Assessors should make consistency checks of their prior probabilities. Various checks are possible. For the beta case, one can assess probabilities of intervals as in the general case and compare these probabilities with areas under the beta density. If they disagree, then some adjustment in a and b may be in order. Another consistency check in the beta case is to assess the probability of success on the second trial, assuming that the first trial results in a *failure*. This probability is

$$r^- = \frac{a}{a + b + 1}.$$

Finding r^- using this expression with a and b as found from the formulas for r and r^+ can be compared with the assessed value of r^-. If the two disagree, the a and b should be adjusted until they do agree.

Another consistency check in the beta case involves $a + b$. Probabilities change less radically when $a + b$ is large. The updating rule for beta densities says that when s successes and f failures are observed, the new density is beta(a+s,b+f). So the new sum of the parameters is the old sum plus n, the sample size. This gives an interpretation for $a + b$ as a 'prior sample size.' Suppose the assessor had experienced a number of observations deemed to be roughly exchangeable with the current observations. Then the assessor might use these observations in setting a to be the number of prior successes and b to be the number of prior failures. So $a + b$ is a measure of reliability in the prior distribution that compares directly with the number of observations in the experiment. This might provide a good primary means of assessing a and b, or at least $a + b$. However, people cannot remember their prior observations very well. Also, prior observations should seldom if ever be regarded as exchangeable with current observations. A possible solution is to discount prior observations as compared with current observations, taking a and b smaller than they would be otherwise. This reflects greater openmindedness and may be a reasonable tack.

Predictive probabilities

A probability distribution of parameters allows for calculating probabilities of responses of *future* observations. Consider a patient with a particular set of characteristics. How will that patient respond to therapy A? The patient's response is unknown. Like all unknowns in the Bayesian approach, it has a probability distribution. Because it refers to future observations, it is predictive. Predictive probabilities are useful for deciding whether to make the observations.

Number of positive lymph nodes

To continue the example of the previous section, suppose a surgeon removes three axillary lymph nodes from a woman with breast cancer and none test positive. Another doctor questions the surgeon, suggesting that had more been removed, perhaps some would have been positive. Given what we now know, should the surgeon have removed more? Should the patient have another surgery? These are complicated questions that require addressing the purpose and utility of nodal dissection — will therapy be different if the patient is node-positive, and how beneficial would it be? It will not be possible to do justice to this issue here. But the probability can be addressed that, if, for example, ten additional lymph nodes were removed, none would be positive.

Let w be the proportion of the patient's lymph nodes that would test positive. As in the previous section, suppose w has a beta(0.14,4.56) prior density. The posterior density of w is then beta(0.14,7.56). For notational convenience, take a and b to be the current values of the beta parameters: $a = 0.14$ and $b = 4.56$. The probability that the next (i.e., fourth) node selected would be negative is $b/(a + b) = 4.56/4.70$. Given that the fourth is negative, the probability that the fifth would also be negative is $(b + 1)/(a + b + 1) = 5.56/5.70$. And so on. So the probability that every one of the next 10 are negative is

$$\frac{b}{a+b}\frac{b+1}{a+b+1}\cdots\frac{b+9}{a+b+9} = \frac{4.56}{4.70}\frac{5.56}{5.70}\cdots\frac{13.56}{13.70}$$
$$= 0.970 \times 0.975 \times \ldots \times 0.990 = 0.84.$$

(The individual factors in this sequence show how the probability of 'negative' increases as additional negative evidence accrues.) So this patient is very likely to continue to be regarded as node-negative, even if an additional 10 nodes are tested. Whether an 84% chance is small enough to recommend more surgery is open to question; the point is that this calculation is an appropriate consideration in such a decision.

Calculations during a phase II trial

Predictive probabilities help in choosing from among possible clinical trial designs. When calculated during the course of a trial, they aid in deciding whether to alter the trial's design — for example, in deciding whether to stop the trial. Consider a phase II trial for evaluating a new agent in the treatment of breast cancer that is newly diagnosed as metastatic. (I have a particular agent in mind.) Now let w be the true rate of response (complete plus partial) in this population. Numerous first-line agents exist. While their response rates are imperfectly known, it is safe to say that some are as large as 30% but that no currently available agents have a response rate much greater than 30%. The probabilities for *w* have been assessed as described in the above section on beta densities and they are well approximated by a beta(2,4) density, shown in figure 5. The mean of this density is 2/(2 + 4) = 1/3, which is the probability of response on the first patient. This is rather large as compared with existing agents, but previous experience with this agent in patients with other types of solid tumors is promising.

Suppose the trial is ongoing and 10 patients have been treated, with 1 of 10 responding and the other 9 not (and so are either stable or have progressed). In deciding whether to continue this trial and include 10 more patients, say, it would be important to know the predictive probabilities for the number of future responders. Using the result of the previous section, the updated distribution of *w* is the beta(2+1,4+9) = beta(3,13) density. The predictive probabilities of the number *k* of successes in the next 10 trials are easy to find:

$$\frac{15!}{2!12!}\,\frac{10!}{k!(10-k)!}\,\frac{(2+k)!(22-k)!}{25!}.$$

These are shown in table 1. Also shown in table 1 are the observed response rates after 20 patients, $(k + 1)/20$. With the addition of 10 patients, *k* of whom are responders, the posterior density of *w* will be beta(k+3,23−k).

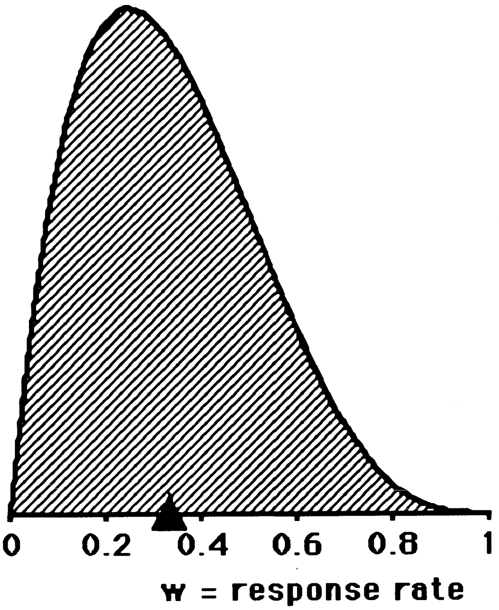


Figure 5. Beta(2,4) prior density for response rate *w* of an experimental therapy for metastatic breast cancer. The mean of 1/3 is shown as a triangle.

The mean of this density is $(k + 3)/26$, which is also shown in table 1. This is another estimate of the response rate, one which includes the prior information as well as the results of the trial. Hence, they are shrunk toward the prior mean of $2/(2 + 4) = 1/3$.

The question of whether the trial should continue will be addressed below. The point here is that the predictive probabilities of these various estimated response rates can be calculated and are relevant for the decision problem.

Calculations during a phase III trial

In May 1984, the Cancer and Leukemia Group B opened a phase III clinical trial for patients with stage III non-small cell lung cancer. The design called for randomizing 240 patients equally to two treatment regimens: radiotherapy alone (RT) and radiotherapy after chemotherapy (RT+CT). Using a truncated O'Brien–Fleming stopping rule, the trial was stopped at the fifth interim analysis in May 1987 after 155 eligible patients had been entered. George et al. (1994) review statistical and other considerations leading to the decision to stop the trial and give Bayesian alternatives to the standard frequentist approaches. Their use of predictive probabilities will be summarized here.

Suppose exponential distributions of the survival times, with λ_1 the death rate on RT and λ_2 the death rate on RT+CT. Then the sufficient statistics are the numbers of patients assigned (n_1 and n_2), the numbers of deaths (d_1 and d_2), and the total time patients had survived on the treatments (T_1 and T_2). These are shown in table 2 for the five interim analyses and also for the final analysis. (Though 155 eligible patients had entered the study by the fifth interim analysis, information was available on only 105 of them.) The

Table 1. Predictive probabilities of the number of responses in next 10 patients given one response in first 10

k	Prob of k	Total #resp.	$\frac{k+1}{20}$	$\frac{k+3}{26}$
0	0.198	1	0.05	0.12
1	0.270	2	0.10	0.15
2	0.231	3	0.15	0.19
3	0.154	4	0.20	0.23
4	0.085	5	0.25	0.27
5	0.040	6	0.30	0.31
6	0.016	7	0.35	0.35
7	0.005	8	0.40	0.38
8	0.001	9	0.45	0.42
9	0.000	10	0.50	0.46
10	0.000	11	0.55	0.50

Table 2. Sufficient statistics for λ_1 and λ_2 assuming exponential model, by analysis time

	On RT			On RT+CT		
Analysis	n_1	d_1	T_1	n_2	d_2	T_2
1st Interim	25	7	122.0	25	3	164.4
2nd Interim	41	12	240.6	38	4	341.1
3rd Interim	41	20	298.5	47	14	432.8
4th Interim	46	24	376.0	49	18	532.7
5th Interim	51	32	441.8	54	24	611.1
As of 1992	77	71	1135.7	78	65	1737.6

Table 3. Various posterior probabilities for log hazard ratio w and the posterior mean survival times at the interim and latest analyses

	Posterior probabilities for w			Mean survival (mos)	
Analysis	$w < 0$	$w < -0.25$	$w < -0.5$	RT	CT+RT
1st Interim	0.976	0.911	0.794	17.9	52.9
2nd Interim	0.997	0.984	0.940	20.1	77.7
3rd Interim	0.987	0.916	0.720	15.6	31.5
4th Interim	0.985	0.895	0.650	16.2	30.2
5th Interim	0.990	0.909	0.637	14.2	25.9
As of 1992	0.999	0.939	0.523	16.2	27.0

notable early difference in the numbers of deaths in the two groups was maintained through the latest analysis, with the benefit in favor of RT+CT.

Consider the log hazard ratio, $w = \ln(\lambda_2/\lambda_1)$, and assume λ_1 and w are independent. A conjugate prior for λ_1 is the gamma: for $\lambda_1 > 0$,

$$f(\lambda \mid la,b) \propto \lambda_1^{a-1}e^{-b\lambda_1}$$

Evidence available in 1984 is consistent with $a = 2$ and $b = 20$. Regarding w, George et al. take the prior distribution when finding the posterior distribution of v to be standard normal and argue that this distribution is open-minded in the sense that the likelihood function dominates the prior distribution when finding the posterior distribution. This is good and bad. It is bad because a small number of observations may seem more persuasive than is appropriate. One should carefully assess the available information on the relative benefits of the treatments, and use an open-minded prior only if there is essentially none.

The current distribution of λ_1 and of w can be found at any time and used to influence the future course of the trial. Table 3 shows various calculations from the current distribution at each of the interim analysis times and also as of 1992.

That the trial stopped early created something of a controversy. In a letter to the editor commenting on the original publication, Souhami et al. [6] concluded: 'In our view, no firm conclusions can be drawn from such a small study presented in this way. It is a great pity that this otherwise excellent trial was not left unanalyzed and allowed to reach a size at which any difference in survival could have been quantified with reasonable certainty.'

Before concluding that a trial should continue, one should ask what will happen if it does. Obviously, this is random. The distribution of various measures that will be available at the time of a future analysis (including *P*-values) can be found from the predictive distribution of the future observations. If this trial had continued beyond the fifth interim analysis, additional information would have come from (1) the 99 survivors who were in the trial (including the 50 patients in the trial but for whom there was no information available) and who would be followed as usual, (2) the additional 85 patients who would be randomized and subsequently followed (assuming the same accrual rate as for the first 155 patients). This information can be simulated using the information available in May 1987 at the fifth interim analysis. Li [7] does this by assuming exponential survival and Qian [8] does it by assuming Weibull survival (showing that the answers are not sensitive to assuming exponential survival).

From the perspective of Souhami et al. [6], a more interesting comparison is whether there would be a different conclusion *today*. For such a comparison, the 99 survivors in the study in 1987 would have been followed. George et al. [9] assume the information to have been available in 1992 and ask how different it would have been had the trial not stopped. Figure 6 shows the predictive distributions of the mean lifetimes on RT and on RT+CT. The triangles show their current values, 16.2 and 27.0 months, as indicated in table 3. These are the means of the predictive distributions. An interesting aspect of the densities shown in this figure is that the variances are small. So the conclusions concerning the survival times on the two treatments would not have been very different had the study accrued 240 patients. Regarding treatment comparison, figure 7 shows the predictive densities of the posterior distribution of log hazard ratio w: $w < 0$, $w < -0.25$, and $w < -0.5$. The first of these is the probability that survival time is longer on RT + CT than on RT alone; this probability would have changed very little from its current value of 0.999 (see table 3).

Finding predictive distributions is relatively straightforward. Using them is more involved. This is the subject of the next section.

Decision problem

This section will indicate how to address two types of decision problems in a Bayesian fashion: deciding whether to stop a clinical trial, and allocating

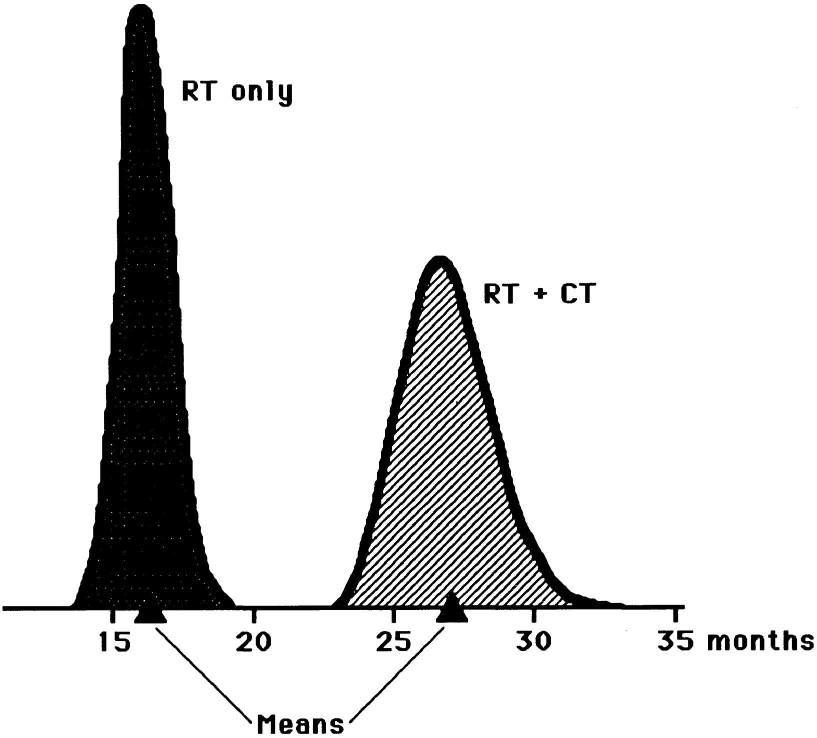


Figure 6. Predictive densities of the posterior mean lifetimes on the two treatment regimens. The triangles indicate the means of the densities, which are also the current values from the fifth interim analysis row of table 3.

patients to therapy to achieve an overall measure of successful health care delivery. For other examples, see [10].

Whether to stop a clinical trial depends on the available information (from the trial and otherwise), given as the current probability distribution of any unknowns. The utility of stopping can be evaluated by weighing the utilities of the various consequences of the status quo by their probabilities. Continuing the trial has the additional randomness provided by the future observations. The utility of continuing can be evaluated by weighing the utilities of the various consequences of future observations by their predictive probabilities. Allocating patients to therapy on the basis of currently available information is a similar problem, one in which predictive probabilities again play a pivotal role.

Three examples of stopping clinical trials are given below. The first is a continuation of the phase II trial example given in an earlier section, and the second is a brief discussion of the non-small cell lung cancer example of the previous section. The third is a randomized vaccine study.

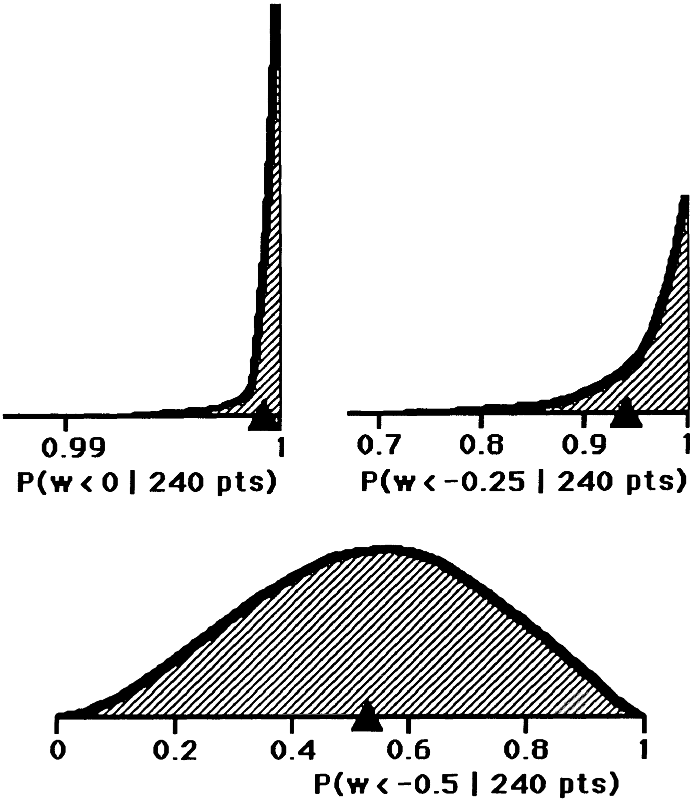


Figure 7. Predictive densities of the posterior probability of various intervals of values of *w*. (These densities are not shown in the same scale.) The triangles indicate the means of the densities, which are also the current values from the fifth interim analysis row of table 3.

Stopping a phase II clinical trial

Deciding whether and when to stop a clinical trial is a common problem, especially since it includes stopping a trial before it starts! In this section I will consider the phase II trial example given in an earlier section, making assumptions about utilities. Recall that one patient responded among the first 10 patients in the trial. I will consider only two possibilities: stopping immediately and adding another 10 patients to the trial. I will consider utilities measured in terms of effective treatment of breast cancer patients. Since there are therapies with 30% response rates and this agent has an estimated 3/16 = 19% response rate, continuing the trial may not be in the best interests of the patients in the trial.

If the trial stops, then I assume this agent will not be investigated further. What would be the resulting impact for the women who have or will have metastatic breast cancer? This is not easy to evaluate, but I will delay the problem by taking it to be 0 and considering other eventualities relative to it. So continuing is appropriate if it has utility greater than 0 and not otherwise.

Now suppose the trial continues to the full 20 patients. The predictive probabilities for the 11 possible numbers of responses in the second 10 patients are given in table 1. I need to specify a utility for each, and I will do so in terms of (my estimates of) the increment in number of responses over the next several years (as compared with the currently available therapies) effected by having a trial in 20 patients with that number of responses. (There is almost certainly a positive relationship between response and survival, but the exact relationship is not clear.) To assess utilities requires addressing several issues. For patients in the trial, how serious is a delay in treatment should it turn out that the experimental agent is not very effective? What other therapies are available, and how effective are they? What are the possibilities that other, perhaps more effective experimental agents will be developed? in what time frame? If the experimental agent turns out to be pomising, what other trials will be necessary before the agent becomes commonly used? As a function of the results from this trial and from other trials, how extensively will the agent be used?

My assessments are shown in table 4. Such assessments should be made by a team of oncologists, pharmacologists, and other experts. But these enable me to demonstrate the method. The first number listed under 'incremental utility' refers to expected difference in responses among next 10 patients in the current trial if they receive this agent as opposed to some other therapy. The second number, the one following the '+', corresponds to patients who present after the trial. This is 0 when the response rate is sufficiently low that, in my estimation, the agent would not be pursued. If

Table 4. Predictive probabilities from table 1 along with the utilities of the various possible number k of responses in the next 10 patients

k	Mean rate	Pred. prob.	Incremental utility	Product
0	0.12	0.198	$-3+0$	-0.59
1	0.15	0.270	$-2+0$	-0.54
2	0.19	0.231	$-1+0$	-0.23
3	0.23	0.154	$0+0$	0
4	0.27	0.085	$1+10$	0.94
5	0.31	0.040	$2+20$	0.87
6	0.35	0.016	$3+60$	0.98
7	0.38	0.005	$4+150$	0.77
8	0.42	0.001	$5+500$	0.63
9	0.46	2E-4	$6+2000$	0.44
10	0.50	2E-5	$7+5000$	0.10
sums:		1.0000		3.37

the response rate is sufficiently high, then later patients are more likely to receive the agent (depending on the results of further clinical trials, an uncertainty I have included in my assessment), and hence the benefit will be greater. Again, the utility is relative to not having this agent available for use.

The (expected) incremental utility of continuing is the average of the fourth column of table 4 — with respect to the predictive probabilities in the third column. The result is 3.37, shown as the sum of products in the fifth column. Units are number of responses, and so this is not a dramatic improvement. But the sum is positive, and so continuing is appropriate.

The sensitivity of the decision (though not its utility) can be judged by varying the utilities in table 4 and also considering prior distributions for w other than the beta(2,4). Such considerations show that the decision to continue the trial is not really close. The sum of the negative products in the fifth column of table 4 is -1.36; the sum of the positives is 4.73. The overall sum would be positive even if the utilities were greatly reduced — with all numbers greater than 10 set equal to 10, say. Only an assessment as extreme as ignoring the patients who present after the trial would make the sum negative. (Because of the evident trade-off, the assessment team should include a medical ethicist.) Also, changing the prior distribution (and hence the predictive probabilities) has little effect on the final result. Prior distributions with more variability (more open-minded) give more probability to smaller values of k but also to larger values of k; because of the asymmetry in the utilities, the net effect is positive. Prior distributions with less variability give less probability to smaller and larger values of k; again the effect is positive, because an agent with a response rate near 30% has positive utility. The only type of prior distribution that results in a negative utility for continuing is one that has a smaller mean and not a very large variance.

Stopping a randomized trial of RT vs. RT+CT in non-small cell lung cancer

The phase III trial discussed above compared RT+CT with RT alone in patients with non-small cell lung cancer. It provides an example of calculating predictive probabilities for survival data. Predictive probabilities aid in deciding whether to stop a trial. I will not address this quantitatively, but my analysis is in line with the decision analysis of the previous example. And the conclusion is the same and is clear: despite the obvious and startling differences between the two therapy regimens and despite the fact that continuing the trial would have had minimal impact on the conclusion, the trial should not have been stopped. My analysis is retrospective. The necessary information was available in 1987, but assembling it would have been difficult at that time. So I suspect that I would have agreed with the investigators in 1987 and come to the wrong conclusion.

In deciding whether to stop a trial, one should assess utilities of the consequences — that is, a cost-benefit analysis. For this, one should consider the impacts on public health, including quality of life of the patients involved. Important considerations are the effects on clinical practice and the impact of study on regulatory officials. A well-designed and well-executed study that is discounted by practitioners is in retrospect not a good study — even if the reasons for discounting are wrong. The non-small cell lung cancer trial showed that adding CT to RT increases lifetime by about 11 months, more than a 60% increase. In 1987 these estimates were about 12 months and more than 80%. Continuing the study means treating half the patients with a regimen that is clearly inferior. Under the circumstances, continuing would be difficult and has questionable ethics.

But how will these and other patients be treated if the study stops? Apparently, most current patients are treated with RT alone. So the trial has not had broad impact, and stopping was unnecessary. Whether it would have had greater impact had it continued is open to question. But it would not have been subject to the specious criticism that it was stopped too early. This would have caused less confusion among practitioners, but may or may not have increased its impact. (The National Cancer Institute has directed a repetition of this trial, which is still ongoing.)

In judging the possible impact of a trial on medical practice, it is important to assess costs and benefits. In the trial discussed here, RT+CT has both monetary costs and toxicity costs (weight loss, infections, vomiting) that partially offset the gain in survival. On the other hand, and 11-month increase in survival may be worth the additional cost, at least for some patients. Cost and benefits should be considered in both the design and analysis of clinical trials.

In general, the question of stopping a clinical trial should depend on whether the interim results will be convincing to the medical community. Whether the results are statistically significant — however that is judged — is relevant only in that it may affect practitioners. Ultimately, one conducts a clinical trial in the hope that the results will influence medical practice. If it were known beforehand that the results of the trial, whatever they might be, would have no influence on practice, most investigators would consider it futile to proceed. The same is true of an ongoing trial at any interim analysis. Investigators should consider the probability that a trial will influence medical practice. Bayesian decision analysis is an appropriate way for making this consideration.

Sequential randomized vaccine efficacy trial

Berry, Wolff, and Sack [11,12] consider a vaccine trial of *hæmophilus influenzæ* type b (HIB). The vaccine was designed to be effective in infants, and the trial led to the licensure of the vaccine for children as young as two months of age. Because Amerind and Eskimo children are at very high risk, the trial was conducted on the Navajo reservation, and the subjects were Navajo children between 2 and 18 months of age. All subjects were vaccinated, with children in the control group received placebo vaccinations.

Berry, Wolff, and Sack [11,12] take the goal of the trial to be minimizing the number of HIB cases among Navajo children over a horizon of 20 years. They assess prior information concerning the rates of HIB and model occurrence among infants as a Poisson process. (Qian [8] generalizes this assumption and models the population as a mixture of two groups: in one the occurrence is Weibull, and the other is not susceptible.) They consider historical information about the regulatory process and assess probabilities of licensure and time required for licensure as a function of the available data. They also assess the possibility that competing vaccines will become available, and the timing of such (for specific assumptions, refer to [12]).

Accumulating information is evaluated ad libitum. At the beginning of the trial and bimonthly thereafter, a decision is made whether to continue the trial or not. The preferred action is the one with smaller expected number of future cases — determined by dynamic programming and exploiting predictive probabilities. (See the following section for a somewhat more detailed discussion of dynamic programming.) Figure 8 is taken from [12] and shows the decision schema. The predetermined maximum length of the trial is N months. At each decision time prior to the Nth month, the trial must be continued or stopped. If it is stopped, a decision is made to seek licensure or not. If licensure is sought, the vaccine will be approved or not.

Table 5 gives the data — the numbers of cases over time in the two treatment groups. Approximately 450 Navajo infants are born each month on the reservations in question. Not all of these were randomized, but they are all at risk for HIB, and we assume they would all receive an approved vaccine. The average accrual rate was 105 children per treatment group per month, so about 240 Navajo infants per month do not participate in the trial. The number of current subjects in each treatment group increased linearly until month 16; when it remained constant at 16*105 = 1680.

The last two columns of table 5 show the results of the dynamic programming. These columns give the expected numbers of future cases of HIB when stopping and when continuing. 'Continuing' assumes that subsequent decisions to stop or continue are optimal. The smaller number is in boldface type. In particular, starting the trial is optimal because 439 is smaller than 608. (For comparison, the expected number of future cases using a fixed rather than sequential design is 553.)

Adaptive allocation in clinical trials

A standard way to compare two therapies is to randomize patients equally to them. For any fixed sample size, this procedure gives maximal information about the difference in their effectivenesses. This in turn will help in treating patients who present once the trial's results become known. In

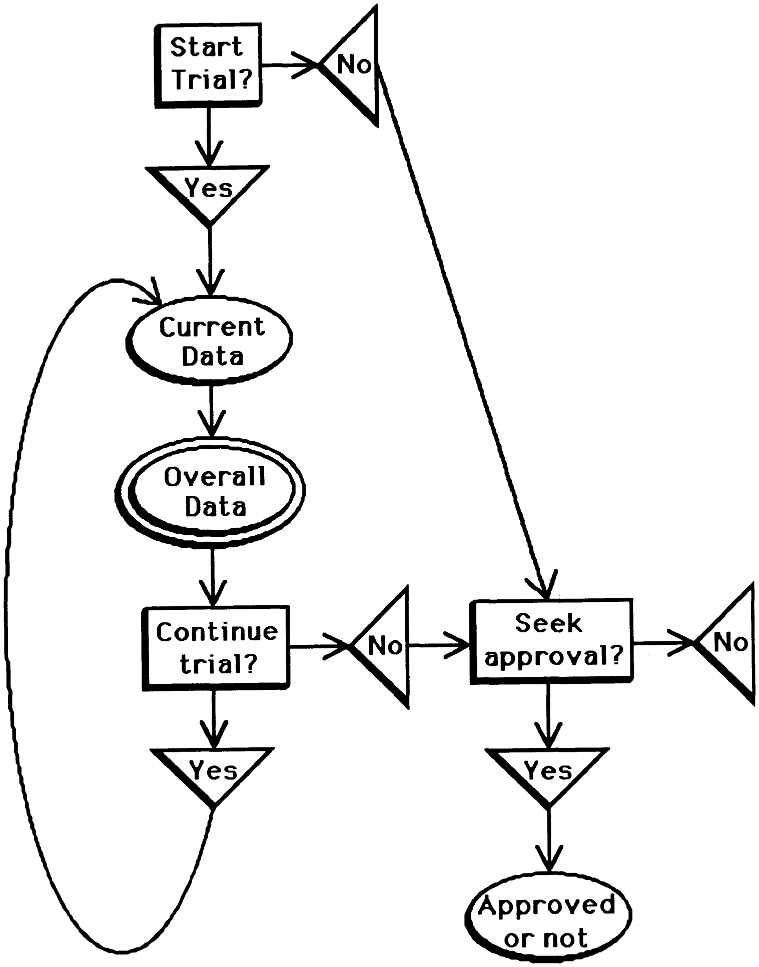


Figure 8. Schema of the decision process. The trial is stopped once the predetermined maximum number N of months is reached. It is also stopped if the expected number of cases over the subject horizon is greater for continuing than for stopping.

designing a trial using a decision approach, one can explicitly consider effective treatment of patients — those in the trial and those not.

Consider a trial in which information accrues relatively quickly concerning each patient's response. (It will be clear that adaptive allocation has no benefit when there are long delays in deciding whether a patient is a responder.) Suppose the objective is to maximize the number of patients who respond, over some patient horizon N. Also, suppose the number of patients in the trial is n. The patients in the trial can be allocated to either

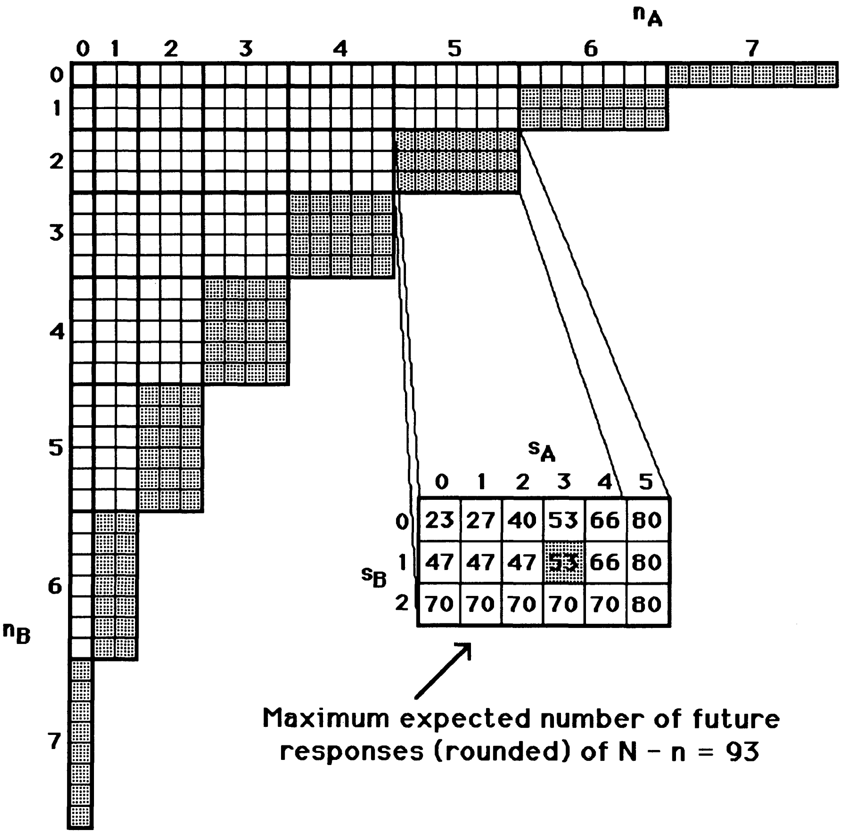
Table 5. Bimonthly results of the trial and expected numbers of future cases (numbers in boldface type indicate the better decision)

	Cumulative cases		Expected number of future cases if	
Month	Vaccine	Placebo	Stop	Continue
0	0	0	608	**439**
2	0	2	807	**505**
4	0	4	954	**498**
6	0	6	1042	**442**
8	0	7	997	**366**
10	0	7	864	**299**
12	0	8	496	**246**
14	0	10	238	**200**
16	1	12	364	**313**
18	1	13	297	**269**
20	1	15	242	**234**
22	1	18	**200**	201
24	1	21	**172**	176
26	1	22	**153**	158

therapy, depending on the accumulating results, and the $N-n$ patients outside the trial will receive the therapy that performs between during the trial. How should the patients in the trial be allocated? To demonstrate the method of finding the optimal procedure, consider an example using $n = 7$ and $N = 100$. A small value of n might be considered in a phase II study, but the reason for choosing such a small n is to be able to draw figures showing the method in a manageable space. In general, increasing n increases the expected number of responses (except that in example below, there is no benefit in increasing an odd n by 1, and so $n = 7$ gives the same expected number of responses as does $n = 8$).

Label one therapy A and the other B. Take the two-population proportion of responses to be w_A and w_B. Take them to be independent, both having a uniform prior distribution on the interval (0,1), which is beta (1,1). Dynamic programming proceeds from the last step of the decision process, which in this case is at the end of the trial. But it requires that all possibilities be considered. Suppose n_A is the number of patients assigned to A and n_B is the number assigned to B. At the end of the trial, $n = n_A + n_B$. One possibility is $n_A = 5$ and $n_B = 2$. Let s_A be the number of responses among the n_A patients assigned to A and s_B the number of responses among the n_B patients assigned to B; s_A takes values 0, 1, . . . , n_A and s_B takes values 0, 1, . . . , n_B. For $n_A = 5$ and $n_B B = 2$, the 18 possible combinations are shown as shaded cells in figure 9.

Consider possibility $n_A = 5$, $n_B = 2$, $s_A = 3$, and $s_B = 1$. The updated mean of w_A is $(3 + 1)/(5 + 2) = 4/7$ and that of w_B is $(1 + 1)/(2 + 2) = 2/4$. Since $4/7 > 2/4$, the remaining $N - n = 93$ patients are assigned to A, with



sB \ sA	0	1	2	3	4	5
0	23	27	40	53	66	80
1	47	47	47	53	66	80
2	70	70	70	70	70	80

Maximum expected number of future responses (rounded) of N - n = 93

Figure 9. Beginning the dynamic program, for each of the cell with $n_A + n_B =$ (cells that are shaded), the expected number of future responses are calculated and entered into this tableau. These entries are shown for those cells with $n_A = 5$ and $n_B = 2$.

an expected future number of successes of 93*4/7 = 53.14. This value (rounded to 53) is shown shaded in figure 9. The other values in the figure and the values for cells with $n_A + n_B = 7$ are calculated similarly. Entering these numbers into the tableau initializes the dynamic program.

Now consider *those cells* for which $n_A + n_B = 6$. Figure 10 shows the one with $n_A = 4$, $n_B = 2$, $s_A = 3$, and $s_B = 1$ The expected number of future responses from this cell is calculated under the two possibilities: use A and use B. If A, then the process moves to one of the two cells shaded in the figure. Referring back to figure 9, the maximal expected number of future responses is 66.43 (with probability 4/(4 + 2)) and 53.14 (with probability 2/(4 + 2)). In the first case, the patient being treated is a responder. So the expected number of future responses when using A is (1 + 66.43)*2/3 + 53.14*1/3 = 62.67. This is compared with the expected number of future responses when using B, and the larger number is entered in the tableau.

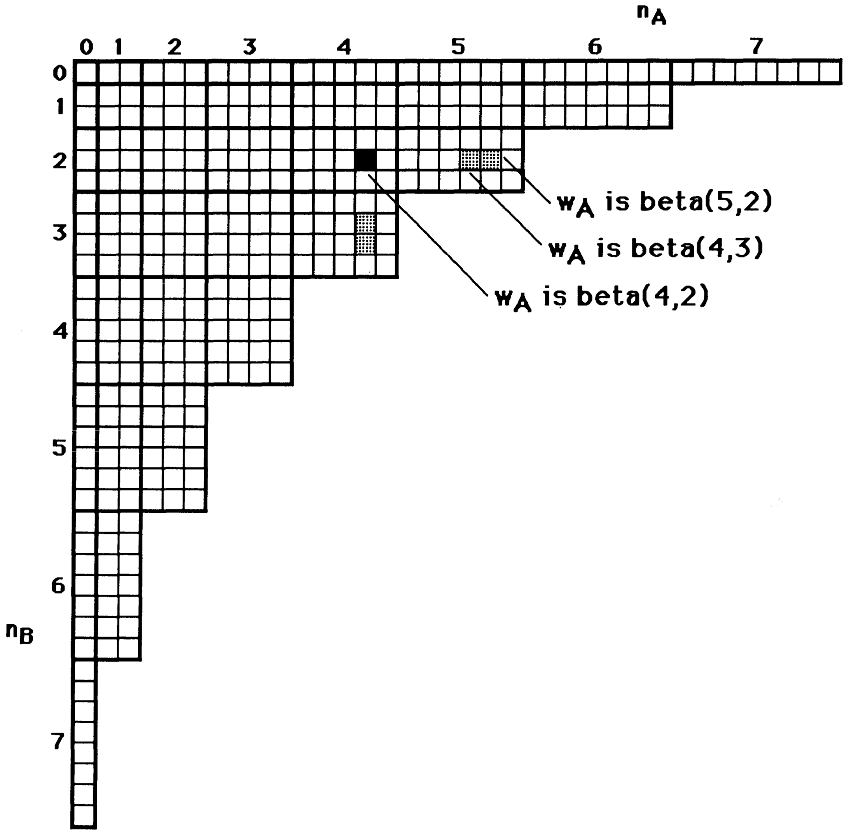


Figure 10. The next steps of the dynamic program are to fill in the tableau for those cells with $n_A + n_B = 6$. The solid cell in the figure has $n_A = 4$, $n_B = 2$, $s_A = 3$, and $s_B = 1$. Using treatment A moves the process to one of the two shaded cells to the right, and using treatment B moves down to one of the two shaded cells. The distribution of w_A for this cell is beta(4,2) and that of w_B is beta(2,2). If treatment A is used, then the distribution of w_A changes to either beta(5,2) (with probability 4/(4 + 2) = 2/3) or beta(4,3) (with probability 1/3), and the distribution of w_B stays the same.

After those cells with $n_A + n_B = 6$ come those with $n_A + n_B = 5$, etc., until we reach $n_A = n_B = 0$. The entry for this cell is 63.05, as indicated in figure 11. This is then the maximal expected number of responses among the $N = 100$ patients. Keeping track of the treatment that gives the larger number of future responses for each cell provides the optimal allocation procedure. This is shown in figure 12.

For comparison, letting $N = n = 100$ gives a maximal expected number of responses of 64.92. (When $N = n$, the decision problem is a classical bandit — see [13]). For a randomized trial with $n = 7$ with 3 patients assigned to A and 4 to B, or vice versa, the expected number of responses in 62.4.

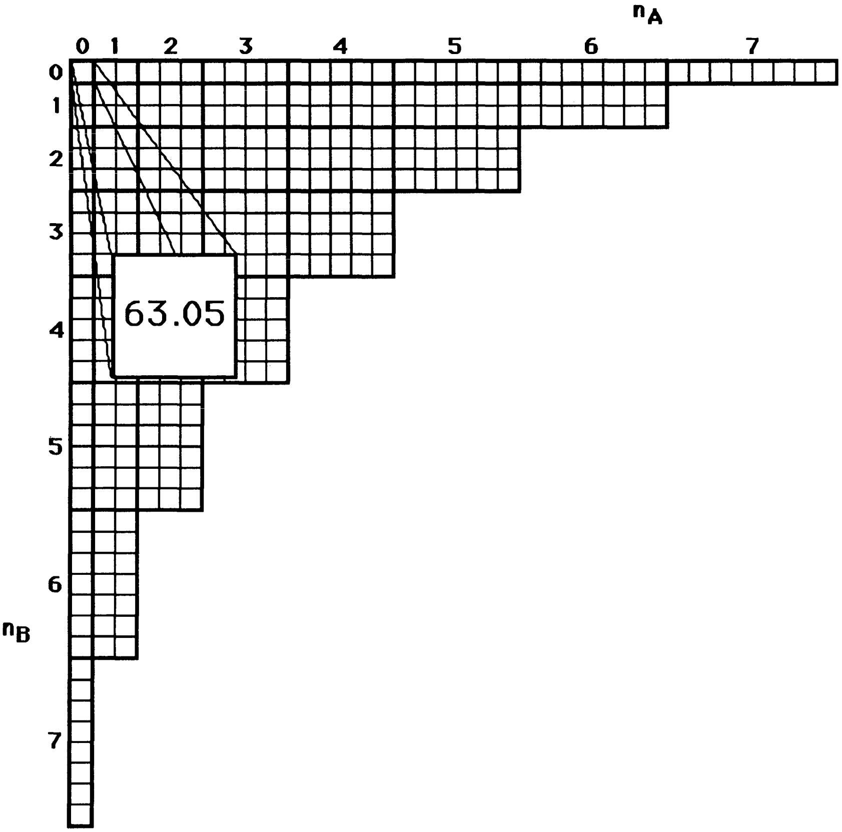


Figure 11. The last step of the dynamic program gives the expected number of responses over the N = 100 patients, namely, 63.05.

Berry and Eick [14] call the optimal Bayesian procedure for uniform prior distributions found above the 'robust Bayes' procedure. They compare it with various adaptive procedures and with a balance randomized controlled trial with sample size n. No procedure can perform better than the robust Bayes procedure, on the average. But they compare procedures for fix w_A and w_B.

For example, suppose N = 10,000 and n = 100. Figure 13 shows the expected number of responses lost as compared with robust Bayes, for two particular procedures. One is RCT, the randomized controlled trial in which patients are assigned equally to A and to B. The other is PW, which is play-the-winner procedure, one of several adaptive procedures by Berry and Eick [14]. Under PW, the first patient is randomized, and then the same treatment is used after a response (play the winner) and the other treatment is used after a nonresponse (switch on a loser). For all three procedures,

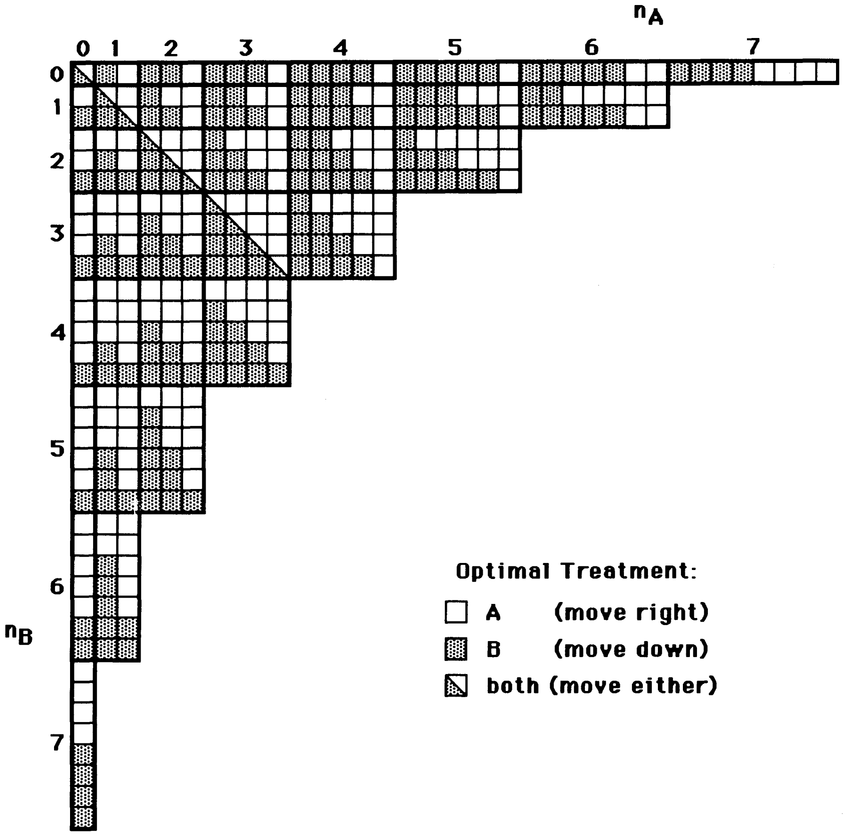


Figure 12. The optimal decisions, depending on the currently available data, as given by n_A, n_B, s_A, and s_B.

following the experimental phase (the first n patients), the better-performing treatment is used exclusively. The figure considers three values of w_B, and w_A varies from 0 to 1. An interesting aspect of the curves is that they stay positive, indicating that robust Bayes is better than both alternatives for given w_A and w_B as well as on the average — hence, 'robust Bayes.'

More generally, RCT loses to robust Bayes for all n, N, w_A, and w_B; but it fares relatively well when patient horizon N is very large. So from a decision-analytic perspective, an RCT is a reasonable choice when a disease or condition is at least moderately common.

A decision approach is an alternative to choosing a significance level, an alternative hypothesis, and a power level. For those interested in power, it may be difficult to evaluate analytically for an adaptive design. However, it is rather easy to evaluate power for any design using simulation.

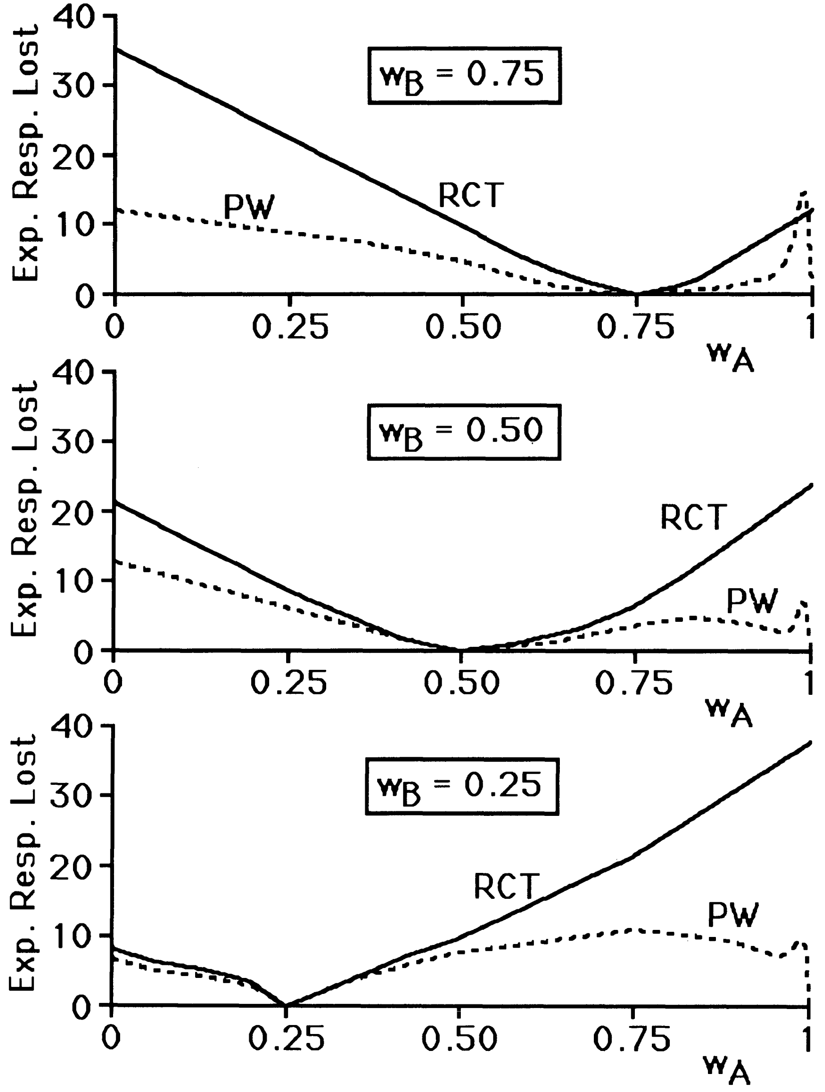


Figure 13. Expected successes lost by RCT and PW as compared with robust Bayes.

Other topics

This chapter provides an introduction to Bayesian methods and decision making in clinical trials. But it barely scratches the surface of some issues, and so far a number of important topics have not been discussed. This section will briefly address some of the latter.

Using posterior probabilities for design

As mentioned above, much of the Bayesian clinical trial literature uses posterior probabilities to play the role usually played by p-values, but does not explicitly address decision issues. Examples include Spiegelhalter, Freedman, and Parmar [3] Carlin et al. [15], and Rosner and Berry [16]. This last paper describes a design of multiple-arm clinical trials in which an arm is dropped if the probability that other arms are better becomes sufficiently large. The motivating example is a trial in metastatic breast cancer patients comparing various infusion schedules and doses of taxol.

Modeling historical information

In interpreting results of clinical trials, historical information is widely regarded as unreliable. However, individual clinicians must place historical information in the context of current information. As I have discussed above, the usual way to incorporate historical information is subjectively, through the prior distribution. Eddy, Hasselblad, and Schachter [17], Berry and Hardwick [18], and Lin [19] introduce methods for discounting historical data through modifications of the likelihood function of the historical data. Such discounting is in part subjective. And it depends partly on the historical data and partly on the current data — especially on the comparability of the two.

Hierarchical models

Hierarchical models are ideally suited for and are commonly used in Bayesian analysis [20,21]. Berry and Stangl [5] provide numerous examples. A typical problem in which a hierarchical model is appropriate is meta-analysis [17,22]. Individual studies are viewed as having unknown characteristics that set them apart from the others. Each study is regarded as having a particular distribution of patient responses for each therapy. Selecting a study means selecting one of these distributions — a random-effects model. If the distribution of the selected study were to be revealed, this would give direct information about how the study distributions are themselves distributed, and there is no hierarchy. But the study's distribution is not revealed. Instead, we get only indirect information about the distribution of study distributions — namely, we get to observe a sample from that distribution by observing the data from that study. We may be interested in a comparison of therapies within a particular study or in the distribution of studies. Difficulties in making computations can be overcome using recent simulation-based calculational advances.

Acknowledgment

This research was supported in part by the National Science Foundation under grant DMS 94-04200 and the Public Health Service under grant RO1 HSO6475-01

References

1. Berry DA (1987). Interim analysis in clinical trials: the role of the likelihood principle. *Am Stat* 41:117–122.
2. Berger JO, Berry DA (1988). The relevance of stopping rules in statistical inference, *Statistical Decision Theory and Related Topics IV*, vol. 1. New York: Springer-Verlag, 29–72.
3. Spiegelhalter DJ, Freedman LS, Parmar MKB (in press). Bayesian approaches to randomized trials (with discussion). *J R Stat Soc A* 157.
4. Berry DA (1993). A case for Bayesianism in clinical trials (with discussion). *Stat Med* 12:1377–1404.
5. Berry DA, Stangl DK (1995). *Bayesian Biostatistics*. New York: Marcel Dekker.
6. Souhami RL, Spiro SG, Cullen M (1991). Chemotherapy and radiation therapy as compared with radiation therapy in stage III non-small cell cancer. *N Engl J Med* 324:1136–1137.
7. Li CC (1994). *Metaanalysis of survival data*. Ph.D. dissertation, Duke University.
8. Qian J (1994). *A Bayesian Weibull Survival Model*. Ph.D. dissertation, Duke University.
9. George SL, Li CC, Berry DA, Green MR. Stopping a clinical trial early: frequentist and Bayesian approaches applied to a CALGB trial in non-small cell lung cancer. *Stat Med* 13:1313–1327.
10. Berry DA (1991). Experimental design for drug development: a Bayesian approach. *J Biopharmaceut Stat* 1:81–101.
11. Berry DA, Wolff MC, Sack D (1992). Public health decision making: a sequential vaccine trial (with discussion). In *Bayesian Statistics*, vol. 4, JM Bernardo, JO Berger, AP Dawid, AFM Smith (eds.). Oxford: Oxford University Press, 79–96.
12. Berry DA, Wolff MC, Sack D. Decision making during a phase III randomized controlled trial. *Controlled Clin Trials* 15:360–378.
13. Berry DA, Fristedt B (1985). *Bandit Problems: Sequential Allocation of Experiments*. London: Chapman-Hall.
14. Berry DA, Eick SG. Adaptive assignment versus balanced randomization in clinical trials: A decision analysis. *Stat Med* 14:231–246.
15. Carlin BP, Chaloner KM, Louis TA, Rhame FS (in press). Elicitation, monitoring, and analysis for an AIDS clinical trial (with discussion). In *Case Studies in Bayesian Statistics*, C Gatsonis, J Hodges, R Kass (eds.).
16. Rosner GL, Berry DA (in press). A Bayesian group sequential design for a multiple arm randomized clinical trial. *Stat Med* 14.
17. Eddy DM, Hasselblad V, Schachter R (1992). *Meta-Analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence*. New York: Academic Press.
18. Berry DA, Hardwick J (1993). Using historical controls in clinical trials: application to ECMO. In *Statistical Decision Theory and Related Topics*, JO Berger, S Gupta (eds.) vol. 5. New York: Springer-Verlag, 141–156.
19. Lin Z (1993). *Statistical Methods for Combining Historical Controls with Clinical Trial Data*. Ph.D. Dissertation, Duke University.
20. Lindley DV, Smith AFM (1972). Bayes estimates for the linear model (with discussion), *J R Stat Assoc B* 34:1–14.

21. Berger JO (1986). *Statistical Decision Theory and Bayesian Analysis*, 2nd edition. New York: Springer-Verlag.
22. DuMouchel W (1989). Bayesian metaanalysis, In *Statistical Methodology in the Pharmaceutical Sciences.* DA Berry (ed.). New York: Marcel Dekker, 509–529.

8. A Bayesian model for evaluating specificity of treatment effects in clinical trials

Richard M. Simon, Dennis O. Dixon, and Boris Freidlin

Introduction

Prognostic factors provide important information for counseling individual patients and for the design of clinical trials. For example, cancer trials of good-prognosis patients may be oriented to reducing toxicity, whereas trials of poor-prognosis patients may focus on improving outcome. Ultimately, one wants to identify factors that permit the selection of appropriate treatments for individual patients. In this way, one could avoid treating many patients with toxic treatments when only a few of those treated would actually derive benefit. Statisticians have often been critical of subset analyses of clinical trials in which one attempts to identify such selectivity of treatment effects. There are several good reasons for this caution. Sometimes subset analyses are merely 'fishing expeditions' attempting to find some subset that appears to give 'positive results' in a clinical trial where there is no overall treatment benefit. Also, because most clinical trials are only sized for detecting overall effects, the power of the trial for detecting true specificity effects is often poor. The probability of claiming specificity by chance alone when none exists is determined by the number of subsets examined and is not limited by the sample size. Consequently, many subset claims are false positives, and the true positives often go undetected.

Several statistical methods have been developed for evaluating whether there is evidence of treatment specificity [1]. Here we shall describe a new method that provides a unified approach to this question in the context of analyzing a clinical trial. The method described here is a Bayesian method based on a relatively simple statistical model. It can be used with time-to-event, binary, or continuous endpoints. Although the method is based on that previously described by Dixon and Simon [2,3], it differs from that method in that it does not require specialized software and is not limited to binary covariates. In this chapter, we first describe a clinical trial for patients with HIV disease that will be used to illustrate the method. Then we provide a standard proportional-hazards-model analysis of the results of this clinical trial. The method itself is described next. We then present the results of the Bayesian analysis and contrast this with the results of the proportional-

Peter F. Thall (ed.), RECENT ADVANCES IN CLINICAL TRIAL DESIGN AND ANALYSIS.
Copyright © 1995 Kluwer Academic Publishers, Boston. All rights reserved. ISBN 978-1-4613-5830-5.

hazards-model analysis. We then conclude with a discussion of extensions of the model and comparisons to the previously published Dixon–Simon model.

The clinical trial

The AIDS Clinical Trials Group (ACTG) is a national clinical trials organization sponsored by the AIDS Division of the National Institute of Allergy and Infectious Diseases. Study ACTG 155 was started in December 1990 to compare three antiretroviral therapy regimens for persons with advanced HIV disease. The regimens were zidovudine (ZDV, 200 mg three times per day), zalcitabine (ddC, 0.75 mg three times per day), and combined ZDV and ddC. The primary objective was to compare the treatment groups in terms of times to occurrence of an AIDS-defining event or death.

Randomizations were stratified on the basis of HIV disease status (symptomatic or asymptomatic), length of previous treatment with ZDV (up to one year or more than one year), and type of prophylaxis for pneumocystis carinii pneumonia (PCP) (local only versus systemic only versus neither or both local and systemic). One thousand and one volunteers had been enrolled from 51 sites by the time accrual stopped in August 1991. Follow-up ended on January 15, 1993. The randomization was weighted 2:2:3 in favor of the combination therapy group.

A detailed presentation of study design and results has been published [10]. Groups were well balanced with respect to pretreatment level of CD4+ T-cells as well as all stratification factors. Investigators intended from the start to examine results in subsets of patients defined by these four characteristics. Figure 1 shows the Kaplan–Meier estimates of the distribution of time without progression or death for the three treatment groups. The logrank test of the homogeneity of these three curves gives a nonsignificant result ($p = 0.28$).

Cox proportional hazards analysis

For the analysis of this clinical trial, we used Cox's proportional hazards (PH) model in the following way. The four covariates were those described above. Each was represented as a binary variable: $x_1 = 0$ for CD4 positive T-cell count less than 100 and 1 otherwise; $x_2 = 0$ for patients without symptoms and 1 otherwise; $x_3 = 0$ for patients not receiving systemic PCP prophylaxis and 1 otherwise; and $x_4 = 0$ for patients who have received ZDV for at least one year and 1 otherwise. Binary representation for CD4 count and PCP prophylaxis was used because the Dixon–Simon method requires binary covariates. The cutoff of 100 for CD4 count was based on this being a commonly used threshold. Two treatment indicators were

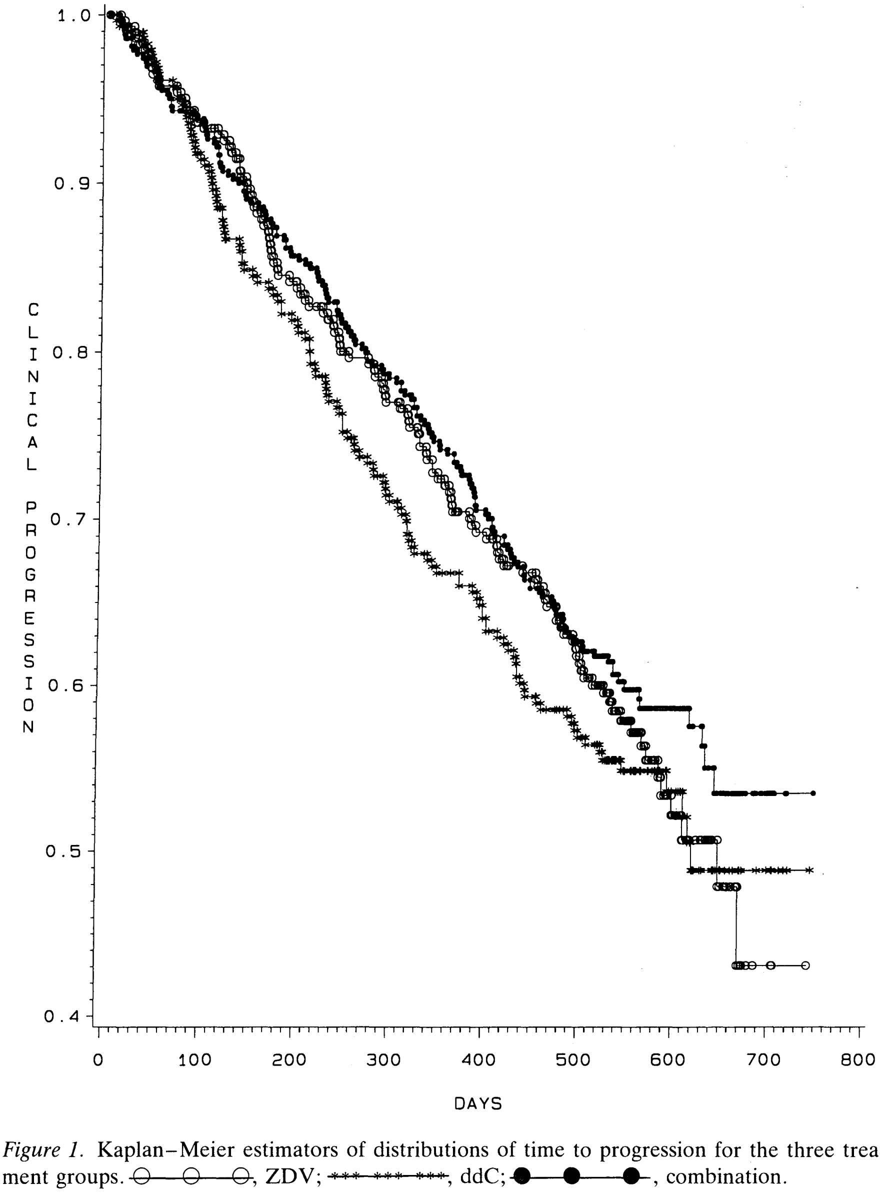


Figure 1. Kaplan–Meier estimators of distributions of time to progression for the three treatment groups. ○—○—○, ZDV; ***—***—***, ddC; ●—●—●, combination.

defined; $z_1 = 1$ for patients receiving ZDV and 0 otherwise; and $z_2 = 1$ for patients receiving ddC and 0 otherwise. Thus patients receiving the combination had both indicators equal to 1.

In the absence of a group receiving neither ZDV nor ddC, 'main effect of ZDV' here represents the contribution of ZDV to the combination, that is, the comparison of combination therapy to ddC monotherapy. Similarly. 'interaction between CD4 count and use of ddC' represents the extent to which the contribution of ddC to the combination depends on CD4 count. It is potentially very misleading to interpret main effect terms when the fitted model also includes interaction terms, as in any regression analysis with cross-product terms. The hazard function of the model used was

$$\lambda(t) = \lambda_0(t)\exp(\alpha_1 z_1 + \alpha_2 z_2 + \beta x + \gamma_1 z_1 x + \gamma_2 z_2 x) \qquad (1)$$

where x is a vector of the four covariates, α_1 and α_2 are regression coefficients corresponding to the main effects of the treatments, β is a vector of regression coefficients corresponding to the main effects of the covariates, and γ_1 and γ_2 are vectors corresponding to the interactions between the covariates and the treatments. Thus $\beta x = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ and likewise $\gamma_1 z_1 x$ and $\gamma_2 z_2 x$ are each the sum of four terms. $\lambda(t)$ and $\lambda_0(t)$ represent the hazard function for AIDS-free survival and baseline hazard at time t.

Table 1 shows the results of this analysis. The only regression coefficients that are statistically significant at the 0.05 two-sided level are the main effect of CD4 cell count (x_1) and the interaction between CD4 count and use of ddC ($z_2 x_1$). We performed a likelihood ratio test of the global hypothesis that all treatments by covariate interactions are zero. The value of the test statistic is 10.937, which has a chi-squared distribution with eight degrees of freedom under the null hypothesis. This yields a significance level greater than 0.25, and hence the hypothesis of homogeneity would not be rejected.

The reduced model without interaction terms is shown in table 2. For this model, the main effect of CD4 count (x_1) is highly significant, the main effect of HIV symptoms (x_2) is highly significant, and the main effect of ZDV (z_1) is also significant.

Table 1. Proportional hazard regression: Foll model

Variable	Parameter estmate	Standard error	Statistical significance
z_1	−0.243	0.338	0.47
z_2	−0.358	0.340	0.29
x_1	−1.091	0.327	0.0009
x_2	0.312	0.424	0.46
x_3	−0.078	0.306	0.80
x_4	−0.606	0.351	0.08
$z_1 * x_1$	0.041	0.268	0.88
$z_1 * x_2$	−0.061	0.341	0.86
$z_1 * x_3$	−0.077	0.242	0.75
$z_1 * x_4$	0.198	0.268	0.46
$z_2 * x_1$	−0.543	0.259	0.036
$z_2 * x_2$	0.241	0.342	0.48
$z_2 * x_3$	0.166	0.246	0.50
$z_2 * x_4$	0.471	0.286	0.10

Table 2. Proportional hazards regression: main effects model

Variable	Parameter estmate	Standard error	Statistical significance
z_1	−0.256	0.120	0.032
z_2	−0.154	0.121	0.20
x_1	−1.431	0.108	0.0001
x_2	0.436	0.141	0.002
x_3	0.003	0.101	0.98
x_4	−0.123	0.114	0.28

Table 3. Proportional hazards regression: reduced model

Variable	Parameter estmate	Standard error	Statistical significance
z_1	−0.261	0.119	0.029
z_2	−0.047	0.149	0.75
x_1	−1.038	0.187	0.0001
x_2	0.445	0.141	0.0016
x_3	0.009	0.101	0.93
x_4	−0.124	0.114	0.28
$z_2 * x_1$	−0.577	0.228	0.012

Table 3 shows the result of the PH regression analysis when the CD4 by ddC interaction is retained in the model. Here we obtain significant main effects of CD4 count, HIV symptoms, and the combination versus ddC contrast, as well as a significant interaction between CD4 count and the combination versus ZDV contrast. The results of both reduced models must be interpreted with caution, however, because they are models selected based on the data. Hence, the regression coefficients and significance levels may be distorted. Tables 1 to 3 are the types of results that are often shown for PH model analysis.

Because of multiple comparison issues, dependence of the regression coefficients on variable selection, and the difficulty of interpreting regression coefficients for models containing interactions, the conclusion to be reached from these analyses is somewhat ambiguous. If we accept the results of table 2, then the conclusion seems to be that the combination is superior to ddC but not superior to ZDV, and there is no statistically significant evidence of treatment effect specificity. Table 1, on the other hand, appears to indicate that the only treatment difference is one between the combination and ZDV, and that difference depends on the CD4 cell count. The benefit of the combination over ZDV is greater for patients with CD4 cell counts above 100. If we accept the results of table 3, however, then we would conclude that the combination is superior to ddC and that there is also evidence that the relative benefit of the combination over ZDV depends on the CD4 cell count.

The Bayesian model

Dixon and Simon [2] introduced a Bayesian model for the analysis of clinical trials with binary covariates. Their approach applies equally to linear models or to the linear combination of covariates used in many nonlinear models, such as logistic models or proportional hazards models. Consider the following proportional hazards model:

$$\lambda(t) = \lambda_0(t)\exp(\alpha z + \beta x + \gamma zx), \tag{2}$$

where x denotes a vector of binary (0,1) covariates, β is a vector of regression coefficients corresponding to these covariates, z is a binary (0,1) treatment indicator variable, α represents the main effect of treatment, γ is a vector of regression coefficients corresponding to the treatment by covariate interactions, and $\lambda_0(t)$ represents the baseline hazard function.

Dixon and Simon assumed that

$$\gamma \sim N(0, \xi^2 I) \tag{3}$$

This incorporates an assumption of exchangeability of interaction effects, in that it assumes that no interactions are a priori more likely than any others and that interactions in one direction are no more likely than those in the opposite direction. Because of this exchangeability assumption, all covariates are required to be of the same scale, e.g., binary.

Dixon and Simon used flat priors for the main effects α and β but introduced a modified Jeffreys hyper-prior for the variance component ξ^2. A hyper-prior is a prior distribution on a parameter of a lower-level prior. The Jeffreys' prior distributions have two properties that are often desirable. They provide little information about the parameter relative to that provided by the experiment. A Jeffreys prior for a parameter also defines a Jeffreys prior for any well-defined transformation of the parameter. Using this hyper-prior, Dixon and Simon derived an expression for the posterior density of any linear combination of the parameters $\theta = (\alpha, \beta, \gamma) = (\alpha, \beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_p)$, where p denotes the number of covariates.

In this chapter, we will investigate a Bayesian model for the analysis of equation (1) that does not involve a hyper-prior for ξ^2. Like the original Dixon–Simon model, this approach is applicable to any model, linear or nonlinear, that incorporates the covariates through a linear functional. These includes linear, generalized linear, logistic, and proportional hazards models. The model to be developed will be used to analyze a clinical trial with three treatment arms. Consequently, we extend model (2) to that shown in (1), where z_1, z_2 are binary (0,1) treatment indicators and γ_1, γ_2 are vectors of treatment by covariate interaction effects.

Let θ denote the vector of parameters $(\alpha_1, \alpha_2, \beta, \gamma_1, \gamma_2)$ and let $\hat{\theta}$ denote the maximum partial likelihood estimate of θ obtained in the usual way by fitting Cox's proportional hazards model (1). Then approximately

$$\hat{\theta}|\theta \sim N(\theta, C) \tag{4}$$

where the covariance matrix C will be assumed known. We shall assume that θ has a normal prior distribution with zero mean vector and covariance matrix D; that is,

$$\theta \sim N(0, D). \tag{5}$$

The assumption of zero prior mean for the main effects will be of no consequence because we will use flat independent priors for these parameters by letting the corresponding diagonal elements of D approach infinity. It follows from Lindley and Smith [4] that the posterior distribution of θ is also normal:

$$\theta|\hat{\theta} \sim N(Bb, B), \tag{6}$$

where

$$B^{-1} = C^{-1} + D^{-1} \tag{7}$$

and

$$b = C^{-1}\hat{\theta}. \tag{8}$$

If independent priors are assumed, if flat priors are used for the main effects, and if the vectors γ_1 and γ_2 are exchangeable, then D^{-1} becomes a diagonal matrix with main diagonal equal to

$$(0_{p+2}, 1/d_1, 1/d_2, \ldots, 1/d_p, 1/d_1, 1/d_2, \ldots, 1/d_p) \tag{9}$$

where, if there are p covariates, 0_{p+2} is a vector of $p + 2$ zeros corresponding to the reciprocals of the prior variances of the main effects, and d_i denotes the prior variance of the ith component of both γ_1 and γ_2.

Given any linear combination $\eta = a\theta$ of the parameters, the posterior distribution of η is also normal:

$$\eta|\theta \sim N(aBb, aBa), \tag{10}$$

where b and B are given above.

Specification of priors

In order to compute the posterior distributions, we must specify the prior variances $d_1, d_2, \ldots, d_p$ of the interaction terms. Consider the effect of the combination relative to treatment 2 alone for a patient with a covariate vector with all components zero. This effect is α_1. Let $x^{(i)}$ be a covariate vector with all components zero except the ith. Let the ith component have value x_i. For binary covariates, $x_i = 1$. The effect of the combination relative to treatment 2 for a patient with covariate vector $x^{(i)}$ is $\alpha_1 + \gamma_{1i} x_i$. Let Δ denote the smallest treatment difference that is considered clinically significant. Define

$$\pi = \Pr[\alpha_1 + \gamma_{1i} x_i \leqslant \Delta | \alpha_1 = 0]. \tag{11}$$

Thus π is the probability of a clinically significant treatment effect for a patient with all covariates zero except for the ith component, given that there is no treatment effect for a similar patient with all covariates zero. A $\leqslant$ symbol is used in equation in (11) because a negative log-hazard corresponds to a beneficial treatment effect in the proportional hazards model. Since the prior distributions of the main effects and interactions are assumed independent, we obtain

$$\begin{aligned} \pi &= \Pr[\gamma_{1i}\, x_i \leqslant \Delta] \\ &= \Phi(\Delta / x_{\mathrm{i}} \sqrt{d_i}) \end{aligned} \tag{12}$$

For binary covariates, $x_i = 1$, and by specifying π and Δ, we can solve equation (12) for d_i. For example, $\Delta = \log(0.5) = -0.69$ corresponds to a halving of the hazard of failure. A skeptical prior for interaction might correspond to specifying a priori that the probability of halving the hazard in one elementary subset, given that the effect is zero in an 'adjacent' subset, is 0.05. This gives $d_i = 0.177$. In the results to be presented here, we have used $\Delta = -0.69$ and have evaluated results for $\pi = 0.05$ and 0.01. We have used the same prior variance for all interaction terms, although this is not necessary. For a continuous covariate, x_i could be taken as the inter-quartile range. Because the model is invariant to adding a constant to a covariate for all patients, the prior may be specified in this manner.

The computations required for using this approach to analysis are thus rather straightforward and require no specialized software. The maximum likelihood estimates $\hat{\theta}$ and covariance matrix C are obtained from a standard Cox proportional-hazards-model analysis. The quantities π and Δ are specified, and then equation (12) is back-solved for d_i. This is repeated for different covariates (i) if different prior variances are desired. These values (or value) of d_i define the diagonal matrix D using equation (9); the components of the main diagonal of D corresponding to the main effects are set equal to zero. Equations (7) and (8) are then used to compute the matrix B and the vector b. The posterior distribution of any linear combination of parameters $a\theta$ has a normal distribution, with mean and variance specified by equation (10). The analysis then consists of selecting the linear combinations of interest and computing the corresponding posterior statistics, as in tables 4 to 6 below.

Linear combinations and qualitative interactions

Three types of linear combinations are of interest. In one type, the vector a contains a single 1, and all the other components are 0. This is used for computing the posterior distribution of a particular parameter. The second type of linear combination is for evaluating the posterior distribution of a treatment contrast within an elementary subset of patients. An elementary subset is a subset defined by the simultaneous specification of all four covariates. For example, to evaluate the posterior distribution of the log hazard ratio of failure for the combination versus ddC for patients with CD4 count greater than 100, no HIV symptoms, receiving systemic PCP prophylaxis who have received ZDV for at least one year, then the linear combination is $a = (1,0,\ 0,0,0,0,\ 1,0,1,0,\ 0,0,0,0)$. This is because for such a patient the covariates are $x = (1,0,1,0)$ and the treatment indicators are (1,1) for the combination and (0,1) for ddC alone. There are 2^4 elementary subsets.

We are also interested in evaluating treatment contrasts for subsets determined by each covariate separately — for example, patients with CD4 count greater than 100. Such quantities are not uniquely determined without specifying the values of the other covariates or the distribution of those values. For example, let w_i denote the proportion of cases with $x_i = 1$ and $x_1 = 1$, of those with $x_1 = 1$, for $i = 2,3,4$. Then the average treatment effect of the combination versus ddC alone for cases with CD4 > 100 is taken as $a\theta$ with $a = (1,0,\ 0,0,0,0,\ 1,w_2,w_3,w_4,\ 0,0,0,0)$. This type of linear combination is also used to evaluate treatment contrasts for the patient sample as a whole. The average treatment effect of the combination versus ddC alone overall is taken as $a\theta$ with $a = (1,0,\ 0,0,0,0,\ w_1,w_2,w_3,w_4,\ 0,0,0,0)$, where (w_1,w_2,w_3,w_4) are the average values of the covariates for the sample overall.

Two treatments exhibit a qualitative interaction over a class of subsets if one treatment is preferable for some of the subsets and the other treatment is preferable for other of the subsets. Peto [5] has argued that only qualitative interactions are important because the usual quantitative interactions are scale dependent and there is no reason to expect that treatment effects should be exactly the same for different subsets. Gail and Simon [6] and Piantadosi and Gail [7] have developed significance tests of the hypothesis that there is no qualitative interaction for disjoint subsets. Russek-Cohen and Simon [8] have developed such tests for multiway classifications. Here we shall show how to calculate the probability that a qualitative interaction does or does not exist for a specified treatment contrast and class of subsets. We shall derive this for the case of two subsets, but the results generalize directly to any number of subsets.

Let $\eta_1 = a_1\theta$ and $\eta_2 = a_2\theta$ denote linear combinations that represent the same treatment contrast for two different subsets. The subsets may be of any type, either disjoint elementary subsets or composite subsets, each determined by the level of a single covariate (e.g., individuals with CD4 >

100). A qualitative interaction for the treatment contrast over these two subsets is said to exist if the linear combinations are not of the same sign. The probability that both linear combinations are positive is easily computed because η_1 and η_2 are jointly normal. The means and variances of η_1 and η_2 are given by equation (10) and the covariance is a_1Ba_2. Consequently, the probability that both linear combinations are positive can be computed fairly easily. Similar calculations provide the probability that they are both negative as well as the probability that one is positive and the other is negative.

The approach described above can also be used to calculate the probability that the combination is better than both single agents either overall or for a particular subset of patients. To do this, we define the two linear combinations to represent the contrast of the combination treatment versus ZDV and the combination versus ddC for the same group of patients. For example, to determine whether the combination is better than both single agents on the average for the overall population, we use the linear combinations (1,0, 0,0,0,0, 0.56,0.83,0.41,0.27, 0,0,0,0) and (0,1, 0,0,0,0, 0,0,0,0, 0.56,0.83, 0.41,0.27).

Results

Table 4 shows results of the Bayesian analysis for the overall group of patients studied and for subsets determined by the level of a single covariate. Since there are four binary covariates, there are eight such subsets. Three treatment contrasts are defined in table 4 for each set of patients: the combination versus ZDV alone, the combination versus ddC alone, and ddC versus ZDV. Each treatment contrast for each set of patients is defined by a linear combination η of the model parameters. Table 4 shows the mean and standard deviation of the posterior distribution of η as well as the posterior probability that $\eta > 0$. The latter corresponds to inferiority of the combination compared to single agents and to inferiority of ddC compared to ZDV.

The first row of numbers in table 4 indicates that the combination achieves a lower hazard rate than either single agent. The posterior probabilities that these linear combinations are positive are 0.03 and 0.04, respectively, for ZDV and ddC. That is, the posterior probabilities that the combination is better than ZDV or ddC overall are 0.97 and 0.96, respectively. The average reduction in log hazard is approximately 0.24 for each contrast. Although not shown in the table, we also computed the posterior probability that the combination is better than both single agents. This probability is 0.883. There is no evidence that there is a difference in efficacy between single-agent ddC and single-agent ZDV for patients overall. The mean difference is −0.005, and the probability that the difference is positive is 0.49.

The remaining rows of table 4 show the results for subsets determined by the levels of individual covariates. For the combination versus ddC, it is

Table 4. Treatment effects overall and for subsets determined by one covariate ($\pi = 0.05$)

		Combination vs. ZDV			Combination vs. ddC			ddC vs. ZDV		
Subset	n	Mean	σ	Signif[a]	Mean	σ	Signif	Mean	σ	Signf
All patients	991	−0.242	0.130	0.03	−0.237	0.134	0.04	−0.005	0.14	0.49
cd4 < 100	310	−0.004	0.145	0.49	−0.279	0.135	0.02	0.275	0.157	0.96
cd4 > 100	681	−0.433	0.185	0.01	−0.204	0.198	0.15	−0.229	0.206	0.13
No symptoms	172	−0.374	0.254	0.07	−0.193	0.257	0.23	−0.181	0.292	0.27
Symptoms	819	−0.214	0.136	0.06	−0.247	0.139	0.04	0.032	0.147	0.58
No systemic PCP proph.	570	−0.310	0.159	0.03	−0.208	0.163	0.10	−0.102	0.173	0.28
Systemic PCP proph.	421	−0.143	0.176	0.21	−0.280	0.174	0.05	0.137	0.196	0.76
Prior ZDV > 1 yr	726	−0.335	0.142	0.01	−0.263	0.148	0.04	−0.072	0.153	0.32
Prior ZDV < 1 yr	265	0.013	0.220	0.52	−0.167	0.208	0.21	0.180	0.248	0.76

[a] Posterior probability that linear combination is positive.

seen that the mean difference is relatively consistent across subsets. There is also substantial consistency for superiority of the combination compared to ZDV, but there is little evidence for superiority of the combination over ZDV for patients with initial CD4 counts below 100 or for patients who either have not received ZDV or have received it for less than one year. There is little evidence for the superiority of either single agent compared to the other single agent for any of these sets of patients.

Table 4 was computed using $\pi = 0.05$ in the notation of the above section about specification of priors. Table 5 presents the same results using $\pi = 0.01$, which represents a somewhat more skeptical a priori view of the likelihood that there exist major treatment by subset interactions. The results in these two tables are quite similar, however.

Table 6 shows summary results for the comparison of the combination versus the single agents in each of the 16 elementary subsets. The covariate values are coded 'y' for yes and 'n' for no in an attempt to simplify reading this complex table. The symbol 'sx' denotes symptoms. The table shows the number of patients in each subset, the posterior mean of the linear combination representing the treatment comparison, and the 95% highest posterior density (HPD) interval. The latter is simply the posterior mean, plus or minus 1.96 times the posterior standard deviation. These intervals contain the true value of the treatment contrasts with 95% probability.

For the contrasts between the combination and ddC alone, the results are relatively uniform favoring the combination. The most extreme mean values tend to correspond to the smallest subsets (some of which contain very few patients), and this is reflected in the width of the highest posterior density interval.

For the contrasts between the combination and ZDV alone, the results appear less uniform. Evidence for the superiority of the combination is strongest for patients with CD4 counts >100 who have received ZDV for more than one year, but the posterior intervals for most other subsets are wide and consistent with either a uniform effect or with differential effects.

The usual regression coefficients, standard errors and statistical significance values, based on standard frequentist model regression analyses, provide limited information. In our Bayesian analysis, we have emphasized the presentation of posterior distributions of treatment contrasts for subsets of patients or for averages across subsets. In fact, we will not even present the posterior distributions or 'significance' of individual regression coefficients in our model. Although we have rarely seen such presentations, frequentist analyses could present point estimates and confidence intervals for such linear combinations.

Table 7 shows frequentist results for the elementary subsets using the full model of table 1. The symbol 'mle' denotes the 'maximum likelihood estimate' of treatment contrast within the subset computed from the proportional hazards model. In computing confidence intervals, one must decide

Table 5. Treatment effects overall and for subsets determined by one covariate ($\pi = 0.01$)

		Combination vs. ZDV			Combination vs. ddC			ddC vs. ZDV		
Subset	n	Mean	σ	Signif[a]	Mean	σ	Signif	Mean	σ	Signif
All patients	991	−0.229	0.128	0.04	−0.234	0.131	0.04	0.006	0.138	0.52
cd4 < 100	310	−0.036	0.141	0.40	−0.281	0.132	0.02	0.245	0.154	0.94
cd4 > 100	681	−0.384	0.173	0.01	−0.197	0.183	0.14	−0.186	0.197	0.17
No symptoms	172	−0.328	0.224	0.07	−0.196	0.226	0.19	−0.132	0.267	0.31
Symptoms	819	−0.208	0.133	0.06	−0.243	0.134	0.04	0.034	0.144	0.59
No systemic PCP proph.	570	−0.286	0.152	0.03	−0.208	0.155	0.09	−0.078	0.168	0.32
Systemic PCP proph.	421	−0.145	0.166	0.19	−0.273	0.163	0.05	0.128	0.188	0.75
Prior ZDV > 1 yr	726	−0.300	0.137	0.01	−0.250	0.142	0.04	−0.050	0.150	0.37
Prior ZDV < 1 yr	265	−0.033	0.200	0.44	−0.193	0.191	0.16	0.160	0.232	0.75

[a] Posterior probability that linear combination is positive.

Table 6. Posterior means and posterior probability intervals for elementary subsets ($\pi = 0.01$)

					Combination vs. ZDV		Combination vs. ddC	
cd4 > 100	sx	Systemic PCP proph.	zdv < 1 yr	*n*	Mean	HPD interval	Mean	HPD interval
y	n	n	n	39	−0.59	(−1.12, −0.063)	−0.15	(−0.69, 0.40)
y	y	n	n	208	−0.47	(−0.85, −0.083)	−0.20	(−0.60, 0.20)
y	y	y	n	134	−0.35	(−0.76, 0.061)	−0.26	(−0.69, 0.17)
y	y	y	y	43	−0.11	(−0.62, 0.40)	−0.19	(−0.70, 0.31)
y	n	y	n	33	−0.47	(−1.03, 0.082)	−0.21	(−0.77, 0.36)
y	n	n	y	8	−0.35	(−0.96, 0.27)	−0.08	(−0.70, 0.53)
y	n	y	y	8	−0.23	(−0.87, 0.41)	−0.14	(−0.78, 0.50)
y	y	n	y	77	−0.23	(−0.71, 0.26)	−0.14	(−0.62, 0.35)
n	y	n	n	131	−0.13	(−0.48, 0.21)	−0.28	(−0.62, 0.05)
n	y	y	n	119	−0.019	(−0.38, 0.34)	−0.34	(−0.69, 0.006)
n	y	y	y	55	0.23	(−0.24, 0.69)	−0.28	(−0.70, 0.14)
n	y	n	y	52	0.11	(−0.33, 0.55)	−0.22	(−0.64, 0.20)
n	n	y	n	22	−0.14	(−0.65, 0.37)	−0.29	(−0.78, 0.21)
n	n	y	y	7	0.10	(−0.49, 0.70)	−0.23	(−0.78, 0.33)
n	n	n	y	15	−0.013	(−0.58, 0.56)	−0.17	(−0.72, 0.39)
n	n	n	n	40	−0.26	(−0.74, 0.23)	−0.23	(−0.71, 0.25)

Table 7. Maximum likelihood estimates, 95% confidence intervals, and bonferroni adjusted 95% confidence intervals for comparing the combination to ZDV in elementary subsets

cd4 > 100	sx	Systemic PCP proph.	zdv < 1 yr	mle	Unadjusted confidence interval	Adjusted confidence interval
y	n	n	n	−0.90	(−1.63, −0.17)	(−2.13, 0.33)
y	y	n	n	−0.66	(−1.13, −0.19)	(−1.45, 0.13)
y	y	y	n	−0.49	(−1.01, 0.02)	(−1.36, 0.38)
y	y	y	y	−0.02	(−0.70, 0.65)	(−1.16, 1.11)
y	n	y	n	−0.73	(−1.52, 0.05)	(−2.06, 0.59)
y	n	n	y	−0.43	(−1.30, 0.44)	(−1.90, 1.04)
y	n	y	y	−0.26	(−1.19, 0.67)	(−1.83, 1.30)
y	y	n	y	−0.19	(−0.82, 0.44)	(−1.24, 0.87)
n	y	n	n	−0.12	(−0.54, 0.30)	(−0.82, 0.59)
n	y	y	n	0.05	(−0.38, 0.48)	(−0.68, 0.78)
n	y	y	y	0.52	(−0.06, 1.10)	(−0.46, 1.50)
n	y	n	y	0.35	(−0.21, 0.92)	(−0.59, 1.30)
n	n	y	n	−0.19	(−0.90, 0.51)	(−1.37, 0.99)
n	n	y	y	0.28	(−0.56, 1.12)	(−1.13, 1.69)
n	n	n	y	0.11	(−0.68, 0.91)	(−1.23, 1.45)
n	n	n	n	−0.36	(−1.02, 0.31)	(−1.48, 0.76)

how to deal with the multiple comparison problem, since there are numerous subsets of interest. That is, frequentist analyses often attempt to ensure that the confidence intervals presented will simultaneously cover the unknown parameters 95% of the time or that the probability that any type 1 error is made in an experiment is no greater than 5%. Table 7 shows only those elementary subsets and only contrasts between the combination and single agent ZDV, but there are many other contrasts and subsets of interest. There is, of course, an extensive literature on multiple comparison procedures for linear contrasts in analysis of variance problems. In table 7 we show two types of confidence intervals. One column gives intervals unadjusted in any way for multiplicity. The other column gives confidence intervals incorporating a Bonferroni adjustment for the 48 combinations resulting from three treatment contrasts for 16 elementary subsets. Clearly, there are at least this many contrasts of interest.

In a comparism of the corresponding entries in table 7 and the columns of table 6 corresponding to the combination versus ZDV, several points become apparent. The maximum likelihood estimates of the treatment effects are more variable among subsets than the posterior means. The values of the posterior mean are 'shrunken' toward the overall posterior mean. This generally, but not always, implies shrinkage of the mean towards zero. Some of the mle values are quite extreme. Another difference is that the 95% highest posterior density intervals are considerably narrower than the unadjusted confidence intervals. The Bonferroni adjusted confidence intervals are so broad as to be useless.

Table 8 presents results for the same contrasts as shown in table 6 based on the hierarchical Bayesian model of Dixon and Simon. The hierarchical model provides even greater attenuation of subset differences than the normal model described here. It borrows information external to a subset to a much greater degree, as can be seen from the relative narrowness of the widths of the posterior intervals for subsets containing very few patients.

Table 9 shows results analogous to those shown in table 4 when CD4 count is modeled linearly as a continuous variable rather than as a binary indicator of >100 or <100. The prior variance for the interaction effects corresponding to CD4 are specified using equation (12) with $x_i - x_i^{(i)}$ equal to the difference between the average CD4 count for patients with CD4 values greater than 100 and the average CD4 count for patients with CD4 values less than 100. These averages were similar to the 75th and 25th percentiles. We also used $\pi = 0.05$ in computing the values in table 9. The results are generally similar to those in table 4, except that the effect of the combination compared to ZDV overall is somewhat less significant (0.09 instead of 0.03).

Discussion

One of our objectives has been to extend the method of analysis introduced by Dixon and Simon for use with clinical trials having more than two treatment groups. Both of the clinical trials previously used to illustrate this method were multiarm trials, although the analysis was limited to only two of the arms [2,3]. The ability to analyze all arms in one unified model is important for efficiently estimating the main effects of covariates and for providing a consistent interpretation of treatment contrasts. With three treatment groups, there are two independent contrasts, and hence two indicator variables were used for treatment. In general, K-1 indicator variables should be used with K treatment groups. The coding of these indicator variables is not critical. Although the coding determines the interpretation of individual regression coefficients, such interpretations are problematic in any case. Average treatment effects either overall or for subsets can be obtained by appropriate specification of linear combinations for any coding. Since we place a locally uniform prior on the regression coefficients associated with the treatment indicators, a similar locally uniform prior distribution results for all linear combinations of these indicators. Hence, the parameterization is only a matter of convenience. For special designs in which locally uniform priors may not be desired, the parameterization is more important, and other approaches may be needed. This is the case, for example, with factorial designs if a priori one expects small interactions or for dose-response designs if monotone relationships are expected.

For the antiretroviral trial, our analysis suggests that, for averages over all patients, the combination appears more effective than either single agent.

Table 8. Posterior means and 95% highest posterior density (HPD) intervals for elementary subsets

cd4 > 100	sx	Systemic PCP proph.	zdv < 1 yr	*n*	Combination vs. ZDV		Combination vs. ddC	
					Mean	HPD interval	Mean	HPD interval
y	n	n	n	39	−0.27	(−0.89, 0.06)	−0.22	(−0.59, 0.24)
y	y	n	n	208	−0.27	(−0.71, 0.03)	−0.22	(−0.54, 0.12)
y	y	y	n	134	−0.23	(−0.63, 0.08)	−0.25	(−0.59, 0.10)
y	y	y	y	43	−0.16	(−0.54, 0.25)	−0.24	(−0.61, 0.17)
y	n	y	n	33	−0.24	(−0.81, 0.11)	−0.23	(−0.64, 0.22)
y	n	n	y	8	−0.22	(−0.75, 0.18)	−0.22	(−0.62, 0.33)
y	n	y	y	8	−0.19	(−0.69, 0.27)	−0.23	(−0.66, 0.29)
y	y	n	y	77	−0.20	(−0.59, 0.17)	−0.22	(−0.57, 0.20)
n	y	n	n	131	−0.16	(−0.45, 0.14)	−0.26	(−0.55, 0.02)
n	y	y	n	119	−0.12	(−0.41, 0.23)	−0.29	(−0.6, −0.01)
n	y	y	y	55	−0.08	(−0.40, 0.49)	−0.27	(−0.62, 0.06)
n	y	n	y	52	−0.10	(−0.42, 0.38)	−0.25	(−0.58, 0.09)
n	n	y	n	22	−0.16	(−0.54, 0.24)	−0.27	(−0.66, 0.11)
n	n	y	y	7	−0.11	(−0.48, 0.47)	−0.26	(−0.68, 0.18)
n	n	n	y	15	−0.14	(−0.52, 0.36)	−0.24	(−0.63, 0.22)
n	n	n	n	40	−0.20	(−0.60, 0.16)	−0.24	(−0.61, 0.13)

Table 9. Treatment effects overall and for subsets determined by one covariate, CD4 continuous ($\pi = 0.05$)

		Combination vs. ZDV			Combination vs. ddC			ddC vs. ZDV		
Subset	*n*	Mean	σ	Signif[a]	Mean	σ	Signif	Mean	σ	Signif
All patients	991	−0.167	0.125	0.09	−0.232	0.125	0.03	0.064	0.136	0.68
cd4 < 100	310	−0.036	0.020	0.40	−0.237	0.129	0.03	0.221	0.147	0.93
cd4 > 100	681	−0.384	0.030	0.01	−0.247	0.213	0.12	−0.184	0.224	0.21
No symptoms	172	−0.328	0.050	0.07	−0.181	0.306	0.28	−0.206	0.323	0.26
Symptoms	819	−0.208	0.018	0.05	−0.256	0.154	0.05	0.039	0.240	0.59
No systemic PCP proph.	570	−0.286	0.023	0.03	−0.233	0.182	0.10	−0.111	0.189	0.28
Systemic PCP proph.	421	−0.145	0.027	0.19	−0.256	0.187	0.09	0.151	0.207	0.77
Prior ZDV > 1 yr	726	−0.30	0.019	0.01	−0.294	0.160	0.03	−0.099	0.165	0.27
Prior ZDV < 1 yr	265	−0.033	0.040	0.43	−0.103	0.236	0.33	0.255	0.273	0.82

[a] Posterior probability that linear combination is possible.

The posterior probability that the average log hazard for the combination is lower than that for ddC is 0.96; for the combination versus ZDV, the figure is 0.97. For the comparison of the combination to ddC, the effect appeared consistent across the subsets of patients. For the comparison of the combination to ZDV, the effect was only conclusive for patients with CD4 counts greater than 100 who had received ZDV for more than one year. The results were less conclusive in the other subsets, but in none was there an indication that ZDV was more effective than the combination. The conclusion that the combination is superior to ZDV overall is somewhat weakened when CD4 is modeled as a continuous variable rather than as a binary indicator, as seen in table 9.

Has this analysis produced greater insight or different conclusions than the Cox proportional hazards analyses described above? This is primarily for the reader to decide. The usual presentation of the full model, as shown in table 1, provides little information and invites possibly erroneous conclusions. One is tempted to conclude that there are no main effects of either ZDV or ddC because neither regression coefficient approaches statistical significance. One is cautioned from this interpretation by the nominally significant interaction between the ddC effect and CD4 cell count, but there are no interactions that approach significance involving the ZDV effect.

Usually, frequentist analyses do not stop with full models retaining numerous nonsignificant variables. Model reduction and variable selection procedures are quite varied and ad hoc, however (e.g., [9]). One approach often used in clinical trials where treatment by subset interactions is not expected is to test the global null hypothesis that all interactions are zero. This provides protection against the possibility that at least one interaction will appear significant by chance. As noted in the earlier discussion of Cox proportional hazards analyses, this approach results in the main effect model shown in table 2. In this model, there is a significant main effect of ZDV that was not apparent in the full model. There is also a significant main effect of symptoms that was not apparent in the full model. Alternatively, one may ignore the multiple comparison issue of eight interaction terms and eliminate all interactions except the one showing nominal significance in the full model. This model, shown in table 3, has significant main effects of ZDV, CD4, and symptoms, as well as the retained ddC by CD4 interaction. One might have obtained a similar model from one of the many types of variable selection regression procedures. But results may have depended on whether forward addition or backward elimination was used, on the nominal significance level cutoffs used for determining whether variables are retained, on rules for whether main effects are permitted to be eliminated if interactions are retained or on whether variables are tested in groups for elimination. Consequently, one can have little confidence in the appropriateness of a model reduced from the full model using variable selection procedures or in the statistical properties of the regression coefficients and covariance matrix of the selected model.

We had two main motivations for investigating the model described here as a potential competitor to the hierarchical model previously described by Dixon and Simon. First, the computations are simpler for this model. Since the posterior distributions are normal, no special software is needed. Secondly, the hierarchical model is limited to binary covariates for which the interaction effects are exchangeable. These restrictions are easily avoided with the model described here. The results (equations (6), (7), (8), and (10)) do not depend on these assumptions. The diagonal form of the D matrix given in equation (9) depends only on the exchangeability of the interaction vectors for the two treatments, not on exchangeability of the components corresponding to different covariates. The price of this generality is, however, the need to specify priors for all the interaction effects not in an exchangeable set. We have assumed that these effects have zero mean, but that assumption is also not inherent in the approach. As we have illustrated for the CD4 variable, this approach to calibration is also applicable to continuous covariates. Hence, the model described here is readily applied to a wide variety of experiments.

Evaluating treatment effects for heterogeneous populations of patients is a complex endeavor that requires a variety of good tools. Issues of specificity of effects in patient subsets are of increased importance for several reasons, however. First is the development of molecular and genetic characterizations of the differences in disease characteristics among patients. There is increased expectation that these covariates will be important treatment-selection factors, and there is increased emphasis on evaluating such interactions in clinical trials. Second is the increased emphasis on evaluating whether there are gender or minority group differences in treatment effects. We believe that the model examined here may be found useful in other clinical trials.

Acknowledgments

The authors thank the AIDS Clinical Trials Group ACTG 155 investigators and study team who conducted the clinical trial that produced the data set used in our analyses. The AIDS Clinical Trials Group is sponsored by the National Institute of Allergy and Infectious Diseases.

References

1. Simon R (1988). Statistical tools for subset analysis in clinical trials. In *Recent Results in Cancer Research*, Vol III, M Baum, R Kay, H Scheurlen (eds.). New York: Springer Verlag, 55–66.
2. Dixon DO, Simon R (1991). Bayesian subset analysis. *Biometrics* 47:871–882.
3. Dixon DO, Simon R (1992). Rayesian subset analysis in a colorectal cancer clinical trial. *Stat Med* 11:13–22.

4. Lindley DV, Smith AFM (1972). Bayes estimates for the linear model (with discussion). *J R Stat Soc B* 34:1–41.
5. Peto R (1982). Statistical aspects of cancer trials. In *Treatment of Cancer*, KE Halnan (ed.). London: Chapman and Hall, 867–871.
6. Gail M, Simon R (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 41:361–372.
7. Piantadosi S, Gail MH (1993). A comparison of the power of two tests for qualitative interactions. *Stat Med* 12:1239–1248.
8. Russek-Cohen E, Simon R (1993). Qualitative interactions in multifactor studies. *Biometries* 49:467–477.
9. Miller AJ (1990). *Subset Selection in Regression*. London: Chapman and Hall.
10. Fischl MA, Stanley K, Collier AC, Arduino JM, Stein DS, Feinberg JE, Allan JD, Goldsmith JC, Powderly WG, NIAID AIDS Clinical Trials Group (1995). Combination and monotherapy with Zidovudine and Zalcitabine in patients with advanced HIV disease. *Annals of Internal Medicine* 122:24–32.

9. The exact analysis of contingency tables in medical research

Cyrus R. Mehta

Introduction

Modern statistical methods rely heavily on nonparametric techniques for comparing two or more populations. These techniques generate p-values without making any distributional assumptions about the populations being compared. However they rely on asymptotic theory that is valid only if the sample sizes are reasonably large and well balanced across the populations. For small, sparse, skewed, or heavily tied data, the asymptotic theory may not be valid. See Agresti and Yang [1] for some empirical results, and Read and Cressie [2] for a more theoretical discussion.

One way to make valid statistical inferences in the presence of small, sparse, or imbalanced data is to compute exact p-values, based on the permutational distribution of the test statistic. This approach was first proposed by R.A. Fisher [3] and has been used extensively for the single 2 × 2 contingency table. In the past, exact tests were rarely attempted for tables of higher dimension than 2 × 2, primarily because of the formidable computing problem involved in their execution. As we shall see below, these computations are orders of magnitude more difficult that any others previously encountered in statistical inference. Two developments over the past 10 years have removed this obstacle. First, the easy availability of immense quantities of computing power in homes and offices has revolutionized our thinking about what is computationally affordable. Second, many new, fast, and efficient algorithms for exact permutational inference have recently been published. Thus computations that would previously have taken several hours or even days to carry out now take only a few minutes. It only remained to incorporate these algorithms into friendly, well-documented statistical packages. Now this step also has been accomplished. In this chapter, we present a unified framework for exact nonparametric inference, anchored in the permutation principle. We demonstrate that exact statistical inference for a very broad class of nonparametric problems can be accomplished by permuting the entries in

Reprinted from *Statistical Methods in Medical Research*, Vol. 3, No. 2, 1994 with kind permission from Edward Arnold (Publisher) Ltd., London.

Peter F. Thall (ed.), RECENT ADVANCES IN CLINICAL TRIAL DESIGN AND ANALYSIS.
Copyright © 1995 Kluwer Academic Publishers, Boston. All rights reserved. ISBN 978-1-4613-5830-5.

a contingency table subject to fixed margins. Exact and Monte Carlo algorithms for solving these permutation problems are referenced but not described. We then apply these algorithms to several data sets in the form of unordered, singly ordered, and doubly ordered contingency tables. Both exact and asymptotic p-values are computed for these data so that one may assess the accuracy of the asymptotic methods. Finally, we discuss the availability of software to implement the algorithms. Readers primarily interested in applications rather than in the theory behind the permutational principle may skip the next section, and go directly to the section on the analysis of data sets.

Nonparametrics and the permutation principle

For a broad class of statistical tests, the data can be represented in the form of the $r \times c$ contingency table $\boldsymbol{x}$ displayed below:

Rows	Col_1	Col_2	...	Col_c	Row_Total
Row_1	x_{11}	x_{12}	...	x_{1c}	m_1
Row_2	x_{21}	x_{22}	...	x_{2c}	m_2
⋮	⋮	⋮	...	⋮	⋮
Row_r	x_{r1}	x_{r2}	...	x_{rc}	m_r
Col_Tot	n_1	n_2	...	n_c	N

The entry in each cell of this $r \times c$ table is the number of subjects falling in the corresponding row and column classifications. The row and column classifications may be based on either *nominal* or *quantitative* variables. Nominal variables take values that cannot be positioned in any natural order. An example of a nominal variable is profession — medicine, law, business. In some statistical packages, nominal variables are also referred to as *class* variables, or *unordered* variables. Quantitative variables take values that can be ordered in a natural way. An example of a quantitative variable is drug dose — low, medium, high. Quantitative variables may, of course, assume numerical values as well (for example, the number of cigarettes smoked per day).

Unconditional sampling distributions

The exact probability distribution of $\boldsymbol{x}$ depends on the sampling scheme that was used to generate $\mathbf{x}$. When both the row and column classifications are categorical, Agresti [4] lists three sampling schemes that could give rise to $\boldsymbol{x}$;

full multinomial sampling, product multinomial sampling, and Poisson sampling. Under all three schemes the probability distribution of $\boldsymbol{x}$ contains unknown parameters relating to the individual cells of the $r \times c$ table.

Under full multinomial sampling, a total of N items are sampled independently, and x_{ij} of them are classified as belonging to row-category i and column-category j, each with probability π_{ij}. Thus the probability of observing the table $\boldsymbol{x}$ is

$$\Pr(\boldsymbol{x}) = \prod_{i=1}^{r}\prod_{j=1}^{c}\frac{N!\,\pi_{ij}^{x_{ij}}}{x_{ij}!}. \tag{1}$$

Full multinomial sampling might arise, for example, if one were to sample N hospital patients and classify them according to their race (White, Black, Other) and their major medical insurance (Blue Cross, HMO, Other). One would be interested in testing the null hypothesis that race and insurance plan were independent. Formally, let $\pi_{i\cdot}$ be the marginal probability of falling in row-category i, and $\pi_{\cdot j}$ be the marginal probability of falling in column-category j. The null hypothesis assumes that $\pi_{ij} = \pi_{i.}\pi_{.j.}$

Under product multinomial sampling, a predetermined number, m_i, of items are sampled independently from population i, and x_{ij} of them are classified as falling into category j. Let π_{ij} be the conditional probability that an item will fall into category j given that it was sampled from population i. Thus the probability of observing table $\boldsymbol{x}$ is

$$\Pr(\boldsymbol{x}) = \prod_{i=1}^{r}\frac{m_i!\prod_{j=1}^{c}\pi_{ij}^{x_{ij}}}{\prod_{j=1}^{c}x_{ij}!}. \tag{2}$$

Product multinomial sampling might arise, for example, if r drug therapies were being tested in a clinical trial, m_i patients were treated with drug i, and each patient fell into one of c possible categories of response. One would be interested in testing the null hypothesis that the probability of falling into response category j was the same for all i, i.e., the drugs are all equivalent in terms of response. Formally, let π_{ij} be the probability that an individual treated with drug i manifests the response j. The null hypothesis assumes that $\pi_{ij} = \pi_j$, for all $j = 1, 2, \ldots c$, independent of i.

Under Poisson sampling, cell (i, j) of the contingency table accumulates events at a Poisson rate of $N\pi ij$, so the probability of observing table $\boldsymbol{x}$ is

$$\Pr(\boldsymbol{x}) = \prod_{i=1}^{r}\prod_{j=1}^{c}\frac{(N\pi ij)^{x_{ij}}e^{-N\pi_{ij}}}{x_{ij}!}. \tag{3}$$

Poisson sampling might arise, for example, if the entry in cell (i, j) represented the number of induced abortions in district i in year j. One would be interested in testing the null hypothesis that the abortion rate did not change from year to year within a district. Formally, the null hypothesis would assume that the Poisson parameter $\pi_{ij} = \pi_{i.}\pi_{.j}$, where $\pi_{i.}$ is the marginal rate for district i and $\pi_{.j}$ is the marginal rate for year j.

Notice that the above probability distributions for $\boldsymbol{x}$ depend on a total of rc unknown parameters, π_{ij}, $(i = 1, 2, \ldots r)$, $(j = 1, 2, \ldots c)$. Since statistical inference is based on the distribution of $\boldsymbol{x}$ under the null hypothesis of independence of row and column classifications, the number of unknown parameters is reduced (π_{ij} being replaced by $\pi_{i.}\pi_{.j}$ *or* π_j depending on the sampling scheme) but not eliminated. Unknown nuisance parameters still remain in equations (1) to (3), even after assuming that the null hypothesis is true. Asymptotic inference relies on estimating these unknown parameters by maximum likelihood and related methods. But in exact inference we eliminate nuisance parameters by conditioning on their sufficient statistics. This is discussed next.

Exact conditional sampling distributions

The key to exact nonparametric inference is eliminating all nuisance parameters from the probability distribution of $\boldsymbol{x}$. This is accomplished by restricting the sample space to the set of all $r \times c$ contingency tables that have the same marginal sums as the observed table $\boldsymbol{x}$. Specifically, define the reference set

$$\Gamma = \left\{\boldsymbol{y}\colon \boldsymbol{y} \text{ is } r \times c;\ \sum_{j=1}^{c} y_{ij} = m_i; \right. \tag{4}$$

$$\left. \sum_{i=1}^{r} y_{ij} = n_j; \text{ for all } i, j\right\}.$$

Then one can show that, under the null hypothesis of no row and column interaction, the probability of observing any $\boldsymbol{y} \in \Gamma$ is

$$\Pr(\boldsymbol{y}\,|\,\boldsymbol{y} \in \Gamma) \equiv P(\boldsymbol{y}) = \frac{\Pi_{j=1}^{c} n_j!\, \Pi_{i=1}^{r} m_i!}{N!\, \Pi_{j=1}^{c} \Pi_{i=1}^{r} y_{ij}!}. \tag{5}$$

Equation (5), which is free of all unknown parameters, holds for categorical data whether the sampling scheme used to generate $\boldsymbol{x}$ is full multinomial, product multinomial, or Poisson [5].

Since equation (5) contains no unknown parameters, exact inference is possible. However, the nuisance parameters were eliminated by conditioning on the margins of the observed contingency table. Now these margins were not fixed when the data were gathered. Thus it is reasonable to question the appropriateness of fixing them for purposes of inference. The justification for conditioning at inference time on margins that were not naturally fixed at data sampling time has a long history. R.A. Fisher [3] first proposed this idea for exact inference on a single 2 × 2 contingency table. At various times since then, prominent statisticians have commented on this approach. The two reasons most cited for conditioning are *convenience* and *ancillarity*.

Convenience. The margins of the contingency table do not contain any information about the hypothesis under test. Since they are the sufficient statistics for the nuisance parameters, conditioning affords a convenient way to eliminate nuisance parameters and thereby perform exact inference without loss of information.

Ancillarity. The principle underlying hypothesis testing is to compare what was actually observed with what could have been observed in hypothetical repetitions of the original experiment, under the null hypothesis. In these hypothetical repetitions, it is a good idea to keep all experimental conditions unchanged as far as possible. The margins of the contingency table are representative of the nuisance parameters. Fixing them in hypothetical repetitions is the nearest we can get to fixing the values of the nuisance paramters themselves in hypothetical repetitions, since the latter are unknown.

An excellent exposition of the conditional viewpoint is available in Yates [6]. For a theoretical justification, refer to Cox and Hinkeley [7]. Throughout this chapter, we shall adopt the conditional approach. It provides us with a unified way to perform exact inference and thereby compute accurate p-values and confidence intervals, even when the observed $r \times c$ contingency table has small cell counts.

Exact p-value computation

Having assigned an exact probability $P(\boldsymbol{y})$ to each $\boldsymbol{y} \in \Gamma$, the next step is to order each contingency table in Γ by a test statistic or 'discrepancy measure' that quantifies the extent to which that table deviates from the null hypothesis of no row and column interaction. Let us denote the test statistic by a real valued function $D : \Gamma \rightarrow \mathscr{R}$ mapping $r \times c$ tables from Γ onto the real line $\mathscr{R}$. The functional form of D for some important nonparametric tests is specified in the next subsection.

The p-value is defined as the sum of null probabilities of all the tables in Γ that are at least as extreme as the observed table, $\boldsymbol{x}$, with respect to D. In particular, if $\boldsymbol{x}$ is the observed $r \times c$ table, then the exact p-values are obtained by computing

$$p = \sum_{D(\boldsymbol{y}) \geqslant D(\boldsymbol{x})} P(\boldsymbol{y}) = pr\{D(\boldsymbol{y}) \geqslant D(\boldsymbol{x})\}. \tag{6}$$

Classical nonparametric methods rely on the large-sample distribution of D to estimate p. For $r \times c$ tables with large cell counts, it is possible to show that D converges to a chi-square distribution with appropriate degrees of freedom. Thus p is usually estimated by $\tilde{p}$, the chi-square tail area to the right of $D(\boldsymbol{x})$. Modern algorithmic techniques have made it possible to compute p directly instead of relying on $\tilde{p}$, its asymptotic approximation.

This is achieved by powerful recursive algorithms [8] that are capable of generating the actual permutation distribution of D instead of relying on its asymptotic chi-square approximation. We shall see later that p and $\tilde{p}$ can differ considerably for contingency tables with small cell counts.

Choosing the test statistic

As stated previously, the reference set Γ is ordered by the test statistic D. Here we define D for three important classes of problems; general tests on $r \times c$ contingency tables, linear rank tests on $2 \times c$ contingency tables, and odds ratio tests on stratified 2×2 contingency tables.

Tests on $r \times c$ contingency tables. Different test statistics are appropriate for different types of $r \times c$ contingency tables. When both the row and column classifications of the table are nominal, the Fisher, Pearson, and Likelihood ratio statistics are the most appropriate. Tests based on these three statistics are known as omnibus tests, because they are powerful against any general alternative to the null hypothesis.

FISHER. Fisher's exact test orders the tables in Γ in proportion to their hypergeometric probabilities. Specifically, the test statistic for each $\boldsymbol{y} \in \Gamma$ is

$$D(\boldsymbol{y}) = -2\log(\gamma P(\boldsymbol{y})) \tag{7}$$

where

$$\gamma = (2\pi)^{(r-1)(c-1)/2}(N)^{-(rc-1)/2}\prod_{i=1}^{r}(m_i)^{(c-1)/2}\prod_{j=1}^{c}(n_j)^{(r-1)/2}.$$

Fisher [3] originally proposed this test for the single 2×2 contingency table. The idea was extended to tables of higher dimension by Freeman and Halton [9]. Thus, this test is also referred to as the Freeman–Halton test. Asymptotically, under the null hypothesis of row and column independence, the Freeman–Halton statistic has a chi-squared distribution with $(r - 1)(c - 1)$ degrees of freedom [10].

PEARSON. The Pearson test orders the tables in Γ according to their Pearson chi-squared statistics. Thus, for each $\boldsymbol{y} \in \Gamma$, the test statistic is

$$D(\boldsymbol{y}) = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(y_{ij} - m_i n_j/N)^2}{m_i n_j/N}. \tag{8}$$

Asymptotically, under the null hypothesis of row and column independence, the Pearson statistic has a chi-squared distribution with $(r - 1)(c - 1)$ degrees of freedom [4].

LIKELIHOOD RATIO. The Likelihood Ratio test [4] orders the tables in Γ according to the likelihood ratio statistic. Specifically, for each $\boldsymbol{y} \in \Gamma$, the test statistic is

$$D(y) = 2\sum_{i=1}^{r}\sum_{j=1}^{c} y_{ij}\log\left(\frac{y_{ij}}{m_i n_j/N}\right). \tag{9}$$

In many textbooks this statistic is denoted by G^2. Asymptotically, under the null hypothesis of row and column independence, $D(y)$ has a chi-squared distribution with $(r - 1)(c - 1)$ degrees of freedom [4].

KRUSKAL–WALLIS. When there is a natural ordering of the columns of the $r \times c$ table, but the row classifications are based on nominal categories, the appropriate test is the Kruskal–Wallis [4]. One can think of the Kruskal–Wallis test as the nonparametric version of one-way ANOVA. It is used to test the equality of r populations with ordered outcomes. For example, suppose that the r rows represent r different drug therapies, and the c columns represent c distinct ordered responses (such as no response, mild response, moderate response, severe response, etc). The Kruskal–Wallis statistic is more powerful than the Fisher, Pearson, or Likelihood Ratio statistics for detecting shifts in response among the r populations. When there are only two rows in the contingency table, the Kruskal–Wallis test specializes to the Wilcoxon-rank-sum test.

The Kruskal–Wallis test orders the tables in Γ according to the Kruskal–Wallis statistic. Specifically, for each $y \in \Gamma$, the test statistic is

$$D(y) = \frac{12}{N(N+1)[1 - (\lambda/(N^3 - N))]} \sum_{i=1}^{r} [R_i(y) - m_i(N+1)/2]^2/m_i, \tag{10}$$

where λ is the tie correction factor $\Sigma_{j=1}^{c}(n_j^3 - n_j)$, and

$$R_i(y) = y_{i1}(n_1 + 1)/2 + y_{i2}[n_1 + (n_2 + 1)/2] + \ldots + y_{ic}\left[\sum_{j=1}^{c-1} n_j + (n_c + 1)/2\right].$$

Asymptotically, under the null hypothesis that the r populations are the same, $D(y)$ has a chi-squared distribution with $(r - 1)$ degrees of freedom.

When the $r \times c$ contingency table has a natural ordering along both its rows and its columns, the Jonckheere–Terpstra test [11] and the Linear-by-Linear association test [4] have more power than the Kruskal–Wallis test. For example, suppose the r rows represent r distinct drug therapies at progressively increasing doses and the c columns represent c ordered responses. Now one would be interested in detecting alternatives to the null hypothesis in which drugs administered at larger doses produce greater responses than drugs administered at smaller doses. The Jonckheere–Terpstra and Linear-by-Linear association test statistics cater explicitly to such alternatives, for they are better able to pick up departures from the null hypothesis in which the response distribution shifts progressively towards the right as we move down the rows of the contingency table.

JONCKHEERE–TERPSTRA. The tables in Γ are ordered according to the Jonckheere–Terpstra statistic, which is really just a sum of $r(r-1)/2$ Wilcoxon–Mann–Whitney statistics. Specifically, for each $y \in \Gamma$, the test statistic is

$$D(y) = \sum_{i=2}^{r} \sum_{j=1}^{i-1} \sum_{k=1}^{c} [w_{ijk} y_{ik} - m_i(m_i + 1)/2], \tag{11}$$

where the w_{ijk} values are the Wilcoxon scores corresponding to a $2 \times c$ table formed from rows i and j of the full $r \times c$ table. Thus, for $k = 1, \ldots, c$,

$$w_{ijk} = [(y_{i1} + y_{j1}) + \ldots + (y_{i,k-1} + y_{j,k-1}) + (y_{i,k} + y_{j,k} + 1)/2].$$

Under the null hypothesis that the r populations are the same, the Jonckheere–Terpstra statistic has a mean

$$E(D(y)) = (N^2 - \sum_{i=1}^{r} m_i^2)/4 \tag{12}$$

and a variance

$$\begin{aligned}\operatorname{var}(D(y)) = \frac{1}{72}\Bigg[&N(N-1)(2N+5) - \sum_{i=1}^{r} m_i(m_i-1)(2m_i+5) \\ &- \sum_{j=1}^{c} n_j(n_j-1)(2n_j+5)\Bigg] \\ &+ \frac{1}{36N(N-1)(N-2)}\left[\sum_{i=1}^{r} m_i(m_i-1)(m_i-2)\right] \\ &\times \left[\sum_{j=1}^{c} n_j(n_j-1)(n_j-2)\right] \\ &+ \frac{1}{8N(N-1)}\left[\sum_{i=1}^{r} m_i(m_i-1)\right]\left[\sum_{j=1}^{c} n_j(n_j-1)\right].\end{aligned}$$

The asymptotic distribution of

$$Z = \frac{D(y) - E(D(y))}{\sqrt{\operatorname{var}(D(y))}} \tag{13}$$

is normal with mean 0 and variance 1.

LINEAR-BY-LINEAR ASSOCIATION. The tables in Γ are ordered according to the linear rank statistic

$$D(y) = \sum_{i=1}^{r} \sum_{j=1}^{c} u_i v_j y_{ij}, \tag{14}$$

where u_i, $i = 1, 2, \ldots r$, are arbitrary row scores, and v_j, $j = 1, 2, \ldots c$, are arbitrary column scores. Under the null hypothesis of no row by column interaction, the test statistic has mean

$$E(D(y)) = N^{-1}\left(\sum_{i=1}^{r} u_i m_i\right)\left(\sum_{j=1}^{c} v_j n_j\right) \tag{15}$$

and variance

$$\begin{aligned}\operatorname{var}(D(y)) = (N - 1)^{-1}&\left[\sum_i u_{\mathrm{i}}^2 m_i - \frac{(\Sigma_i u_i m_i)^2}{N}\right]\\ &\times\left[\sum_j v_j^2 n_j - \frac{(\Sigma_j v_j n_j)^2}{N}\right].\end{aligned} \tag{16}$$

(Agresti [4], pp. 284 and 303 (problem 8.29), for additional details). The asymptotic distribution of

$$Z = \frac{D(y) - E(D(y))}{\sqrt{\operatorname{var}(D(y))}} \tag{17}$$

is normal with mean 0 and variance 1.

The freedom to select the u_i and v_j scores arbitrarily is a powerful feature of the Linear-by-Linear test [12]. If the u's and v's represent the original raw data, the Linear-by-Linear test is a test of significance for Pearson's correlation coefficient. On the other hand, if the raw data are replaced by ridit or mid-rank scores, we have a test of Spearman's correlation coefficient. For the special case of the $2 \times c$ contingency table, the Linear-by-Linear test statistic yields a rich class of linear rank tests. These are defined next.

Linear Rank Tests. For the special case of the $2 \times c$ contingency table, the Linear-by-Linear association test reduces to the family of linear rank tests

$$D(y) = \sum_{j=1}^{c} v_j y_{1j}. \tag{18}$$

Since we are conditioning on the column sums, it is not necessary to sum over the second row. The scores $\{u_i\}$ have therefore been dropped from the expression for D without any loss of generality.

The mean and variance of D, under the null hypothesis of no row and column interaction, and conditional on $y \in \Gamma$, can be derived from equations (5) and (18). The mean is

$$E(D) = \left(\frac{m_1}{N}\right)\sum_{j=1}^{c} v_j n_j. \tag{19}$$

The variance is

$$\sigma^2 = \left[\frac{m_1 m_2}{N(N - 1)}\right]\sum_{j=1}^{c}\left[v_j - \frac{E(D)}{m_1}\right]^2 n_j. \tag{20}$$

By the Chernoff-Savage theorem [13], the standardized test statistic

$$Z = \frac{D - E(D)}{\sigma} \tag{21}$$

converges in distribution to the standard normal distribution with a mean of 0 and unit variance, under suitable regularity conditions on the scores.

Different choices of scores $\{v_j\}$ yield different linear rank tests. These scores and the conditions under which to use each test are specified below.

WILCOXON SCORES. The Wilcoxon scores

$$v_j = n_1 + \ldots + n_{j-1} + (n_j + 1)/2 \tag{22}$$

are the ranks (midranks in the case of tied observations) of the underlying responses. The Wilcoxon rank-sum test [14] is one of the most popular nonparametric tests for detecting a shift in location between two populations. It has an asymptotic relative efficiency of 95.5%, relative to the t test when the underlying distributions are normal. If there is censoring in the data, the scores defined by equation (22) are replaced by the generalized Wilcoxon–Gehan scores, as discussed in Kalbfleisch and Prentice [15]. In particular, let $a_1, a_2, \ldots a_g$ be the g distinct death times. Let $n_1, n_2, \ldots n_g$ be the corresponding numbers of deaths and $r_1, r_2, \ldots r_g$ be the numbers at risk at these death times. The score assigned to all n_j subjects who die at time a_j is

$$v_{a_j} = 1 - \frac{2}{n_j}\left[\sum_{j=c_{j-1}+1}^{c_j} \prod_{l=1}^{i}\left(\frac{N - l + 1}{N - l + 2}\right)\right], \tag{23}$$

where $c_j = n_1 + n_2 + \ldots + n_j$. For all subjects who are censored between the two death times a_j and a_{j+1}, the corresponding scores are

$$v_{a_{j+}} = 1 - \left.\prod_{l=1}^{c_j}\left(\frac{N - l + 1}{N - l + 2}\right)\right]. \tag{24}$$

Scores for all subjects censored prior to the first failure time are zero. Scores for all subjects censored past the last failure time are computed by equation (23). This convention ensures that the sum of Wilcoxon–Gehan scores over all subjects, and hence the expected value of the Wilcoxon–Gehan statistic, is always zero.

NORMAL SCORES. The scores for the Normal scores (or Van der Waerden) test are the percentiles of the standard normal distribution:

$$v_j = \frac{1}{n_j}\left[\sum_{i=c_{j-1}+1}^{c_j} \Phi^{-1}\left(\frac{i}{N + 1}\right)\right], \tag{25}$$

where $\Phi^{-1}(\alpha)$ is the 100αth percentile of the standard normal distribution. The Normal scores test [14] is an alternative to the Wilcoxon rank-sum test for comparing two populations. It is a nonparametric test with 100%

asymptotic relative efficiency relative to the t test when the underlying distributions are normal with shifted means. If the tails of the distributions are diffuse, however, this test is less powerful than the Wilcoxon.

SAVAGE SCORES. The scores for the Savage test, also known as the exponential scores test, are defined by

$$v_j = \frac{1}{n_j}\left[\sum_{i=c_{j-1}+1}^{c_j}\sum_{l=1}^{i}\left(\frac{1}{N-l+1}\right)\right] - 1. \tag{26}$$

The Savage test is a locally most powerful test [16].

LOGRANK SCORES. Logrank scores are used for censored survival data [15]. They are defined as follows. Let $\{a_j\}$, $\{n_j\}$, $\{r_j\}$, and $\{c_j\}$ be defined as for the Wilcoxon scores. The score assigned to all n_j subjects who die at time a_j is

$$v_{a_j} = \frac{1}{n_j}\left[\sum_{i=c_{j-1}+1}^{c_j}\sum_{l=1}^{i}\frac{1}{N-l+1}\right] - 1. \tag{27}$$

For all subjects who are censored between the two death times a_j and a_{j+1}, the logrank scores are

$$v_{a_{j+}} = \sum_{l=1}^{c_j}\frac{1}{N-l+1}. \tag{28}$$

Scores for all subjects censored prior to the first failure time are zero. Scores for all subjects censored past the last failure time are computed by equation (27). This convention ensures that the sum of logrank scores over all subjects, and hence the expected value of the logrank statistic, is always zero. It can easily be seen that for uncensored data, the logrank scores specialize to the Savage scores defined previously. The Logrank test is a competitor to the Wilcoxon–Gehan test for censored data. It is the optimal test against proportional hazard alternatives. However, for nonproportional hazards with early differences in the hazard rates or crossing hazard functions, the Wilcoxon–Gehan test is more powerful.

TREND. The Trend test [17] uses the equally spaced scores

$$v_j = j. \tag{29}$$

It is also known as the Cochran–Armitage trend test and is a very popular test of a dose–response relationship among c binomial populations, where the jth population is sampled n_j times and each member of the sample is exposed to dose w_j. The probability of a response for each sample is π_j. The null hypothesis is that

$$\pi_1 = \pi_2 = \ldots = \pi_c.$$

The alternative hypothesis is that there is a trend whereby the binomial probabilities, π_j, increase with increasing dose w_j. A variant of the Cochran–Armitage trend test uses the actual doses, w_j, or their logarithms, as the scores instead of replacing them by the equally spaced scores.

Tests on stratified 2 × 2 contingency tables. A very important class of exact nonparametric tests and confidence intervals is defined on data in the form of several 2 × 2 contingency tables. The ith table is of the form

Rows	Col_1	Col_2	Row_Total
Row_1	y_i	x_i	m_i
Row_2	y_i'	x_i'	m_i'
Col_Tot	$N_i - n_i$	n_i	N_i

for $i = 1, 2, \ldots s$. We may regard the two rows of each table as arising from two independent binomial distributions. Specifically, let (x_i, x_i') represent the number of successes in (m_i, m_i') Bernoulli trials, with respective success probabilities (π_i, π_i'). The odds ratio for the ith table is defined as

$$\Psi_i = \left(\frac{\pi_i}{1 - \pi_i}\right) \Big/ \left(\frac{\pi_i'}{1 - \pi_i'}\right). \tag{30}$$

Stratified 2 × 2 contingency tables arise commonly in prospective studies with binary endpoints as well as in retrospective case–control studies. Thus, although we have specified that the two rows of the 2 × 2 table represent two independent binomial distributions, this is just a matter of notational convenience. We could equivalently assume that the two rows represent the disease status and the two columns represent the exposure status in a case–control setting.

We shall be interested in testing the null hypothesis that

$$\Psi_i = \Psi \quad \text{for} \quad i = 1, 2, \ldots s.$$

This is known as the homogeneity test. Next, under the assumption of homogeneity, we shall be interested in estimating the common odds ratio, Ψ. In order to formulate these two problems, we need to extend the notation developed previously for the reference set Γ of $r \times c$ contingency tables with fixed margins. Accordingly, let τ denote a generic set of s 2 × 2 tables. Let τ_0 denote a specific realization of τ. Exact inference, both for testing that the odds ratio across s 2 × 2 tables is constant as well as for estimating the common odds ratio, is based on determining how extreme the observed τ_0 is relative to other τ's that could have been observed in some reference set. Different reference sets are used for testing the homogeneity of odds ratios and for estimating the common odds ratio. Define the reference set

$$\Omega = \left\{ \begin{matrix} \tau\colon & x_i + y_i = m_i;\, x_i' + y_i' = m_i'; \\ & x_i + x_i' = n_i;\, y_i + y_i' = N_i - n_i \end{matrix} \right\}. \tag{31}$$

Also, define the more restricted reference set

$$\Omega_t = \{\tau \in \Omega\colon\ \ x_1 + x_2 + \ldots + x_s = t\}. \tag{32}$$

An exact test for homogeneity of the odds ratios is based on ordering the τ's in Ω_t, while exact inference about the common odds ratio is based on ordering the τ's in Ω. These two exact procedures are discussed next. For completeness, a corresponding asymptotic procedure is also provided next to each exact procedure.

HOMOGENEITY TEST. Zelen [18] developed an exact test for the null hypothesis

$$H_0\colon\ \ \Psi_i = \Psi,\quad i = 1, 2, \ldots s\,.$$

Zelen's test is based on the fact that under H_0 the probability of observing any τ from the conditional reference set Ω_t is a product of hypergeometric probabilities, which does not depend on the nuisance parameter Ψ. Specifically, the conditional probability of obtaining any $\tau \in \Omega_t$ is

$$\Pr(\tau|t) = \frac{\Pi_{i=1}^{s}\binom{m_i}{x_i}\binom{m_i'}{x_i'}\Big/\binom{N_i}{n_i}}{\Sigma_{\tau\in\Omega_t}\Pi_{i=1}^{s}\binom{m_i}{x_i}\binom{m_i'}{x_i'}\Big/\binom{N_i}{n_i}}. \tag{33}$$

In addition to its probabilistic interpretation, equation (33) may be used to order each $\tau \in \Omega_t$ so as to determine how extreme or discrepant the observed τ_0 is under H_0. Thus, $\Pr(\tau|t)$ may also be used as the test statistic for the homogeneity test. Its observed value, $\Pr(\tau_0|t)$, defines the critical region of the exact two-sided p-value. Let

$$\Omega_t^* = \{\tau \in \Omega_t\colon\ \ \Pr(\tau|t) \leq \Pr(\tau_0|t)\}. \tag{34}$$

The p-value for Zelen's test of homogeneity is

$$p = \sum_{\tau\in\Omega_t^*} \Pr(\tau|t). \tag{35}$$

There is no well-accepted large-sample theory for this problem. Breslow and Day [17] propose the statistic

$$\chi^2_{BD} = \sum_{i=1}^{s} \frac{[x_i - A_i(\hat{\Psi})]^2}{\mathrm{var}(X_i|\hat{\Psi})}. \tag{36}$$

where $A_i(\hat{\Psi})$ is the positive root of the quadratic equation

$$\frac{A_i(N_i - m_i - n_i + A_i)}{(m_i - A_i)(n_i - A_i)} = \hat{\Psi}, \tag{37}$$

formed by expressing the *i*th table as

$$\begin{array}{cc} m_i - A_i & A_i \\ N_i - m_i - n_i + A_i & n_i - A_i, \end{array}$$

and equating its empirical odds ratio to the Mantel–Haenszel common odds ratio

$$\hat{\Psi} = \frac{\Sigma_{i=1}^{s} x_i(N_i - m_i - n_i + x_i)/N_i}{\Sigma_{i=1}^{s}(n_i - x_i)(m_i - x_i)/N_i}. \tag{38}$$

The variance of X_i is estimated by

$$\text{var}(X_i|\hat{\Psi}) = \left[\frac{1}{A_i(\hat{\Psi})} + \frac{1}{m_i - A_i(\hat{\Psi})} + \frac{1}{n_i - A_i(\hat{\Psi})} + \frac{1}{N_i - m_i - n_i + A_i(\hat{\Psi})}\right]^{-1}. \tag{39}$$

In large samples, χ^2_{BD} is chi-squared distributed with $s-1$ degrees of freedom, and the p-value for testing H_0 is

$$p_{BD} = \Pr(\chi^2_{BD} \geqslant \chi^2_0), \tag{40}$$

where χ^2_0 is the observed value of χ^2_{BD}. The chi-squared approximation to the χ^2_{BD} statistic is rather poor for skewed or sparse contingency tables.

COMMON ODDS RATIO ESTIMATION. Exact inference about the common odds ratio, Ψ, is based on the fact that the probability of any $\tau \in \Omega$ may be expressed as a product of noncentral hypergeometric probabilities in which Ψ is the only unknown parameter. As shown in Gart [19], this probability is

$$\Pr(\tau) = \frac{\Pi_{i=1}^{s}\binom{m_i}{x_i}\binom{m_i'}{n_i - x_i}\Psi^{x_i}}{\Sigma_{\tau\in\Omega}\Pi_{i=1}^{s}\binom{m_i}{x_i}\binom{m_i'}{n_i - x_i}\Psi^{x_i}}. \tag{41}$$

To make inferences about Ψ, we require the distribution of its sufficient statistic

$$t = x_1 + x_2 + \cdots + x_s. \tag{42}$$

This distribution can be derived from equation (41) as

$$\Pr(T = t|\Psi) = \frac{C_t\Psi^t}{\Sigma_{u=t_{\min}}^{t_{\max}} C_u\Psi^u}, \tag{43}$$

where

$$C_t = \sum_{\tau \in \Omega_t} \prod_{i=1}^{s} \binom{m_i}{x_i} \binom{m_i'}{n_i - x_i}, \tag{44}$$

$$t_{\min} = \sum_{i=1}^{s} \max(0, n_i - m_i), \tag{45}$$

$$t_{\max} = \sum_{i=1}^{s} \min(m_i', n_i). \tag{46}$$

It is straightforward to test the hypothesis $\Psi = \Psi_0$ based on the conditional distribution (equation (43)). The test has critical regions of the form $T \geqslant t$ ($T \leqslant t$) for alternatives of the form $\Psi > \Psi_0$ ($\Psi < \Psi_0$). An exact confidence interval for Ψ may be constructed by inverting this test, as discussed in Cox and Snell [20]. An efficient numerical algorithm for generating the distribution (equation (43)) is given in Mehta, Patel, and Gray [21].

An asymptotic confidence interval for Ψ is usually computed by the Mantel–Haenszel [22] method. The Mantel–Haenszel point estimate, $\hat{\Psi}$, is computed by equation (38). The inference is then based on the large-sample approximation to the distribution of $\log \hat{\Psi}$. This distribution is normal, with mean $\log \Psi$. There has been a great deal of research on the appropriate variance estimator for $\log \hat{\Psi}$. The most satisfactory candidate is the Robins, Breslow, and Greenland (RBG) variance [23]. This variance estimator is known to perform well both when s is small but (m_i, n_i) are large, and when s is large but (m_i, n_i) are small. The RBG variance is

$$\operatorname{var}(\log \hat{\Psi}) = \sum_{i=1}^{s} \left(\frac{a_i c_i}{2c_+^2} + \frac{a_i d_i + b_i c_i}{2c_+ d_+} + \frac{b_i d_i}{2d_+^2} \right) \tag{47}$$

where $a_i = (x_i + y_i')/N_i$, $b_i = (x_i' + y_i)/N_i$, $c_i = (x_i y_i')/N_i$, $d_i = (x_i' y_i)/N_i$, $c_+ = \Sigma_{i=1}^{k} c_i$, and $d_+ = \Sigma_{i=1}^{k} d_i$. A $100(1 - \alpha)\%$ confidence interval for $\log \Psi$ is then

$$CI_{\text{RBG}} = \log \hat{\Psi} \pm z_{\alpha/2} [\operatorname{var}(\log \hat{\Psi})]^{1/2}. \tag{48}$$

Computational issues

Computing equation (6) is a nontrivial task. This is because the size of the reference set grows exponentially, so explicit enumeration of all the tables in Γ soon becomes computationally infeasible. For example, the reference set of all 5 × 6 tables with row sums of (7, 7, 12, 4, 4) and column sums of (4, 5, 6, 5, 7, 7) contains 1.6 billion tables. Yet, the tables in this reference set are all rather sparse and unlikely to yield accurate p-values based on large sample theory. Network algorithms have been developed by Mehta and Patel [8,10,21,24,25] to enumerate the tables in Γ implicitly. This makes it feasible to compute exact p-values for tables with the above margins. A different approach to implicit enumeration is provided by Pagano and

Halvorsen [26], Pagano and Tritchler [27], Baglivo, Olivier, and Pagano [28], and Streitberg and Rohmel [29]. Sometimes a data set is too large even for implicit enumeration, yet it is sufficiently sparse that the asymptotic results are suspect. For such situations, a Monte Carlo estimate and associated 99% confidence interval for the exact p-value may be obtained. In the Monte Carlo method, tables are sampled from Γ in proportion to their hypergeometric probabilities (equation (5)), and a count is kept of all the sampled tables that are more extreme than the observed table. For details, refer to Agresti and Wackerly [30], Patefield [31], and Mehta, Patel, and Senchaudhuri [31].

Analysis of data sets

In this section, we will illustrate the techniques developed in the previous section with some data analysis. Each example will highlight the different conclusions one might draw if an asymptotic analysis were performed instead of an exact analysis.

Unordered contingency tables

In house-to-house surveys in three geographic regions of rural India by Gupta, Mehta, and Pindborg [33], data were obtained on the location of oral lesions. Consider a hypothetical subset of these data in the form of a 9 × 3 contingency table in which each count is the number of patients with oral lesions per site and geographic region.

Site of lesion	Kerala	Gujarat	Andhra
Labial mucosa	0	1	0
Buccal mucosa	8	1	8
Commissure	0	1	0
Gingiva	0	1	0
Hard palate	0	1	0
Soft palate	0	1	0
Tongue	0	1	0
Floor of mouth	1	0	1
Alveolar ridge	1	0	1

The question of interest is whether the distribution of the site of the oral lesion is significantly different in the three geographic regions. The row and column classifications for this 9 × 3 table are clearly unordered, making it an appropriate data set for either the Fisher, Pearson, or Likelihood Ratio tests. The contingency table is so sparse that the usual chi-squared asymptotic distribution with 16 degrees of freedom is not likely to yield accurate p-values. The exact and asymptotic p-values are displayed below.

	Three tests of independence		
Type of inference	Pearson	Fisher	Likelihood Ratio
Value of $D(x)$	22.1	19.72	23.3
Asymptotic p-value	0.1400	0.2331	0.1060
Exact p-value	0.0269	0.0101	0.0356

For each test, the asymptotic p-value was obtained by looking up the tail area to the right of $D(x)$ (displayed on the first line of the table) from a chi-square distribution with 16 degrees of freedom. The exact p-value was obtained by actually permuting the observed 9 × 3 table in all possible ways, subject to fixed margins, and summing the probabilities of permutations y for which $D(y) \geqslant D(x)$. There are striking differences between the exact and asymptotic p-values. The exact analysis suggests that the row and column classifications are highly dependent, but the asymptotic analysis fails to show this.

Singly ordered contingency tables

The tumor regression rates of five chemotherapy regimens — Cytoxan (CTX) alone, Cyclohexyl-chloroethyl nitrosurea (CCNU) alone, Methotrexate (MTX) alone, CTX+MTX, and CTX+CCNU+MTX — were compared in a small clinical trial of non-small cell lung cancer. Tumor regression was measured on a three-point scale: no response, partial response, or complete response. The results are tabulated below.

Chemo	No resp.	Partial resp.	Complete resp.
CTX	2	0	0
CCNU	1	1	0
MTX	3	0	0
CTX+CCNU	2	2	0
CTX+CCNU+MTX	1	1	4

Small pilot studies like this one are frequently conducted as a preliminary to planning a large-scale randomized clinical trial. The columns of the observed 5 × 3 contingency table are ordered by the magnitude of the response. However, the rows of the table do not have any natural ordering, but simply represent five different treatments. For such data, the Kruskal–Wallis test may be used to determine whether or not the five drug regimens are significantly different with respect to their tumor regression rates. The observed value of the Kruskal–Wallis statistic for this table is 8.682. Referring this value to a chi-square distribution with four degrees of freedom yields an asymptotic p-value of 0.0695, which is not significant at the 0.05 level. However, the exact test based on the permutation distribution of equation (10) reveals that the exact p-value is 0.039, which is statistically significant at level 0.05. The small sample size and the presence of ties caused the asymptotic approximation to be nearly twice as large as the exact p-value.

Doubly ordered contingency tables

Dose–response example. Patients were treated with a drug at four dose levels (100 mg, 200 mg, 300 mg, 400 mg) and then monitored for toxicity. The data are tabulated below.

	Drug toxicity				
Drug dose	Mild	Moderate	Severe	Drug death	Row_Score
100 mg	100	1	0	0	u_1
200 mg	18	1	1	0	u_2
300 mg	50	1	1	0	u_3
400 mg	50	1	1	1	u_4
Column_Score	v_1	v_2	v_3	v_4	

Notice that there is a natural ordering along the rows as well as the columns of the above 4 × 4 contingency table. Thus the Jonckheere–Terpstra test and the Linear-by-Linear association test each are appropriate for determining if the increase in drug dose leads to greater toxicity.

We first perform the Jonckheere–Terpstra test. The exact two-sided p-value of 0.1134 closely matches the corresponding asymptotic two-sided p-value of 0.1210, indicating that the dose–response relationship between drug dose and toxicity is not statistically significant. Next we perform the Linear-by-Linear association test, using the equally spaced scores, $u = i$, $v_j = j$, for $i, j = 1, 2, \ldots 4$. Now the exact two-sided p-value is 0.0866 and the corresponding asymptotic two-sided p-value is 0.0812, confirming that the dose–response relationship is at best marginally statistically significant. The Linear-by-Linear association test does give us some added flexibility over the Jonckheere–Terpstra test, however. We are free to choose the row and column scores arbitrarily. Suppose, for instance, that the toxic event 'Drug death' was deemed to be catastrophic, and orders of magnitude more serious than a 'Severe toxicity.' In that case, it might be reasonable to maintain the equally spaced row scores, $u_i = i$, $i = 1, 2, \ldots 4$, but to assign unequally spaced column scores $v_1 = 1$ for 'Mild toxicity,' $v_2 = 2$ for 'Moderate Toxicity,' $v_3 = 3$ for 'Severe toxicity,' and $v_4 = 10{,}000$ for 'Drug Death.' Because of this severe discontinuity in the column scores, the asymptotic theory breaks down. Now the two-sided asymptotic p-value is

0.1604, implying that there is no association between drug dose and toxicity, while the two-sided exact p-value is 0.0372, implying that the dose–response relationship is indeed statistically significant.

Space shuttle Challenger example. Professor Richard Feynman, in his delightful book *What Do You Care What Other People Think?* [34], recounted at great length his experiences as a member of the Presidential Commission formed to determine the cause of the explosion of the space shuttle Challenger in 1986. He suspected that the low temperature at take-off caused the O-rings to fail. On page 137 of his book, he has published the data on temperature versus the number of O-ring incidents on 24 previous space shuttle flights. These data are tabulated below.

O-ring incidents	Temperature (Fahrenheit)								
None	66	67	67	67	68	68	70	70	72
	73	75	76	76	78	79	80	81	
One	57	58	63	70	70				
Two	75								
Three	53								

These data may be represented as a contingency table whose rows are the number of O-ring incidents and whose columns are the temperatures at take-off. Thus both the rows and columns are ordered, and the Jonckhere–Terpstra test is an appropriate one for determining if take-off temperature is correlated with O-ring failures. The exact p-value is 0.0241, while the asymptotic p-value is 0.0262. Both are indicative of a significant association between take-off temperature and O-ring incidents.

The Linear-by-Linear association test may also be used to test the association between temperature and O-ring incidents. Using the number of O-ring incidents as the row scores and the take-off temperature as the column scores, the exact p-value is 0.0272. The corresponding asymptotic p-value is 0.0175. These results confirm the conclusions of the Jonckheere–Terpstra test.

Linear rank tests

A cohort of Hiroshima atomic bomb survivors was followed to determine the relationship between deaths from leukemia during 1950–1970 and estimated radiation dosage from the bombing. Subjects were stratified according to their age at the time of the bombing. Below we tabulate a subset of the data, namely, children in the 0–9 age group exposed to radiation doses ranging from 0 to 99 rads. Cases are subjects who died from leukemia during the follow-up. Controls are subjects who did not die from leukemia during the follow-up.

Survival status	Radiation dose (rads)			
	0	1–9	10–49	50–99
Case	0 (0%)	7 (0.07%)	3 (0.1%)	1 (0.14%)
Control	5015	10,752	2989	694
Total	5015	10,759	2992	695

Two additional dose groups, 100–199 rads and 200+ rads, are excluded from the present analysis. Their inclusion increased the standardized value of the test statistic from 3 to 16, strongly suggesting that their effect on the risk of leukemia is nonlinear and should be considered in a more general model. The full data set is on page 285 of Agresti [4].

In absolute terms, the leukemia death rates are rather low. Only 11 deaths were observed in a cohort of size 19,461, amounting to a death rate of 0.06%. However, the rates increase from 0% in the lowest dose group to 0.14% in the highest. It is therefore interesting to ask whether this increasing trend is real, or merely due to chance fluctuations in the data. Our intuition cannot help much with these extremely low death rates, and we must resort to a formal statistical test of significance.

One way to determine if there is a statistically significant association between leukemia deaths and radiation exposure is to perform the Cochran–Armitage trend test [14]. The test statistic is given by equation (18), with v_j being the midrange of the jth radiation dose. For these data, $v_1 = 0$ rads, $v_2 = 4.5$ rads, $v_3 = 30$ rads, and $v_4 = 75$ rads. Previously the only way to perform this trend test was to assume that the linear rank statistic, D, is normally distributed. Figure 1 displays the true distribution of D. It is not even close to normal. Its distinct values are unequally spaced; the distribution has an unusually long right tail, extending all the way out to $D = 825$ even though $E(D) = 107.6$. In addition, the distribution is multimodal. Not surprisingly, the exact and asymptotic p-values for the Cochran–Armitage trend test differ. The results are tabulated below.

p-values	One-sided	Two-sided
Exact	0.0653	0.0682
Asymptotic	0.0465	0.0929

Stratified 2 × 2 tables

We present two examples in this section, one for a test of homogeneity of odds ratios and one for estimating the common odds ratio.

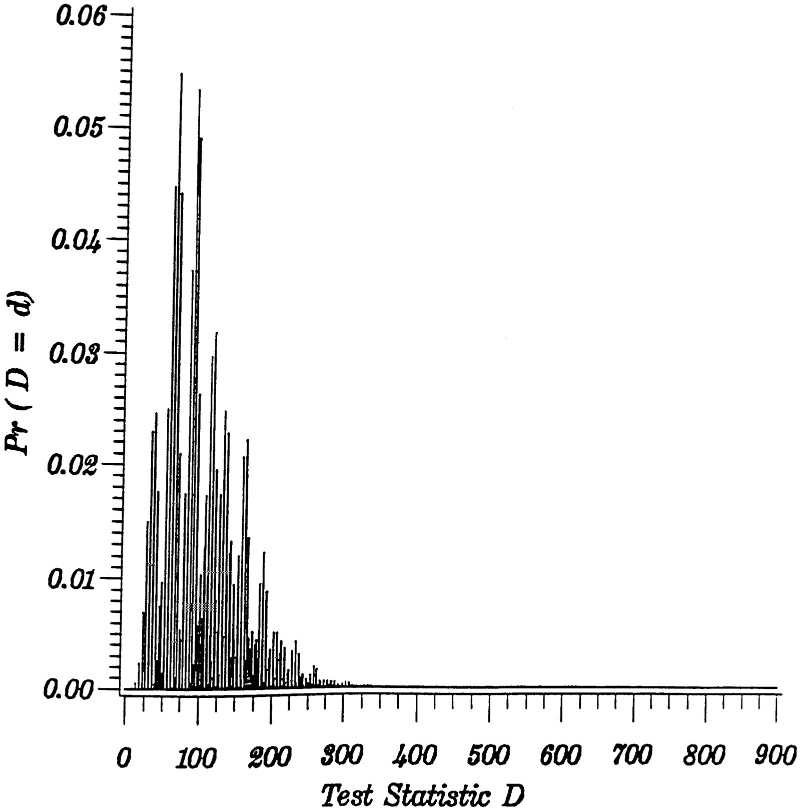


Figure 1. Exact probability density for Hiroshima data.

Homogeneity of odds ratios. The binary response data tabulated below compare a new drug with a control drug at 22 hospital sites. (At the request of the drug company conducting the study, the names of the two agents are not reported here.)

	New drug		Control drug	
Test site	Response	No	Response	No
1	0	15	0	15
2	0	39	6	32
3	1	20	3	18
4	1	14	2	15
5	1	20	2	19
6	0	12	2	10
7	3	49	10	42
8	0	19	2	17

	New drug		Control drug	
Test site	Response	No	Response	No
9	1	14	0	15
10	2	26	2	27
11	0	19	2	18
12	0	12	1	11
13	0	24	5	19
14	2	10	2	11
15	0	14	11	3
16	0	53	4	48
17	0	20	0	20
18	0	21	0	21
19	1	50	1	48
20	0	13	1	13
21	0	13	1	13
22	0	21	0	21

The data can be thought of as 22 2 × 2 contingency tables, one for each site. If you examine the 2 × 2 tables carefully, you notice that site 15 appears to be different from the others. Whereas all the other sites have a low response rate for both the new drug and the control drug, the response rate of the control drug is 79% at site 15. The Homogeneity test can tell you whether the observed difference at site 15 is a real difference or whether it is just a chance fluctuation due to a small sample. Because of the sparseness in the data, the asymptotic (Breslow–Day) statistic might not yield an accurate p-value. The exact (Zelen) test is preferred. The exact p-value is 0.0135. Thus we reject the null hypothesis that there is a common odds ratio across the 22 sites. The data strongly suggest that the odds ratio at site 15 is different from the other odds ratios. The asymptotic (Breslow–Day) p-value is much larger (0.0785) and is only marginally significant.

Estimating the common odds ratio. The court case of Hogan v. Pierce [35] involved the following hiring data, by race.

	Whites		Blacks	
Date of hire	Hired	Not	Hired	Not
7/74	4	16	0	7
8/74	4	13	0	7
9/74	2	13	0	8
4/75	1	17	0	8
5/75	1	17	0	8
10/75	1	29	0	10
11/75	2	29	0	10
2/76	1	30	0	10
3/76	1	30	0	10
11/77	1	33	0	13

The most notable feature of these data is that at each hiring opportunity, not a single black was hired, whereas small numbers of whites were hired. This makes it impossible to use the usual large-sample maximum likelihood or Mantel–Haenszel [22] methods for estimating the odds of being hired for whites relative to blacks. These methods simply fail to converge. Only the exact method provides a valid answer, and it shows that the odds of being hired for a white relative to a black are no lower than 2.3 to 1, with 95% confidence.

Concluding remarks

We have presented the essential idea behind exact nonparametric inference, referenced numerical algorithms and software for its implementation, and shown through several examples that exact inference is a valuable supplement to corresponding asymptotic methods.

The methods described here extend naturally to continuous data. In principle, such data can also be represented as contingency tables, but the columns of these tables will sum to 1. Thus these methods provide a unified approach to handling nonparametric data both for the categorical case and the more traditional continuous case. For example, consider the following two-sample problem involving continuous data. The two groups are 'males' and 'females.' The continuous variable being compared in the two groups is 'monthly income.'

M	M	M	M	F	F	F	F
2010	3100	2555	2095	1990	2122	1875	2550

These data can be represented by the following 2 × 8 contingency table, which may then be permuted in the usual way for exact inference.

Rows	Col_1	Col_2	Col_3	Col_4	Col_5	Col_6	Col_7	Col_8	Row_Total
Male	0	0	1	1	0	0	1	1	4
Female	1	1	0	0	1	1	0	0	4
Col_Tot	1	1	1	1	1	1	1	1	8
Col_Score	1875	1990	2010	2095	2122	2550	2555	3100	

For both representations of the data, the exact two-sided p-value is 0.3429, while the asymptotic two-sided p-value is 0.2965.

In conclusion, exact methods are now an integral part of nonparametric inference. Software support for these methods is available in many standard packages, including SAS. Some of the newer textbooks on nonparametric methods, for example, Sprent [36], devote considerable space to exact methods. Thus one expects that exact methods will replace corresponding asymptotic ones as the standard approach for small, sparse, or unbalanced data sets.

Appendix: software for exact inference

So far as we are aware, there are only five statistical packages meeting commercial standards of reliability and documentation that offer exact inference capabilities beyond the single 2 × 2 contingency table.

EGRET (1989). The EGRET [37] package is available from Statistical-and Epidemiology Research Corporation, 1107 NE 45, Suite 520, Seattle, WA 98105. It offers exact inference for stratified 2 × 2 contingency tables and for the Pearson test for a 2 × c contingency table. Exact inference for the general $r \times c$ problem is not provided.

Epi Info (1989). Epi Info [38] is a series of programs used to create and analyze questionnaires and perform other common epidemiological tasks. One of the statistical capabilities provided by Epi Info is exact inference for the common odds ratio in stratified 2 × 2 contingency tables. It is available from the Division of Surveillance and Epidemiologic Studies, Epidemiology Program Office, Centers for Disease Control, Atlanta, GA 30333.

SAS (1987). SAS [39] is available from the SAS Institute, 100 SAS Campus Drive, Cary, NC 27513. Versions 6 and up offer the exact p-value capability for Fisher's exact test for $r \times c$ tables, but not for any of the other tests described here. A special module, StatXact for SAS (1993), developed by Cytel Software Corporation, Cambridge, MA, extends the exact capabilities of SAS by making it possible to call the StaXact package (described below) from within SAS, read in SAS data sets, and take the results back into SAS so as to make use of SAS's powerful graphics and report generation capabilities.

StatXact (1993). The StatXact [40] package is available from Cytel Software Corporation, 675 Massachusetts Avenue, Cambridge, MA 02139. Version 2 was released in 1991. Version 3 is currently in beta test. It is a complete nonparametrics package with exact tests for one-sample, two-sample, and k-sample problems, measures of association, $r \times c$ contingency tables, stratified 2 × 2 and 2 × c contingency tables, multiple comparisons, exact one- and two-sample Hodges–Lehmann confidence intervals, and exact confidence intervals for odds ratios, risk ratios, and differences in two binomial parameters. It provides software support for standard textbooks on nonparametric statistics such as Lehmann [41], Hollander and Wolfe [11], Gibbons [14], Seigel and Castellan [42], and Sprent [36]. A companion package, LogXact [43], provides exact inference capabilities for logistic regression.

Testimate (1992). The Testimate [44] package is available from IDV, Datenanalyse und Versuchsplanung, Wessobrunner Strasse 6, D-8035 Gauting, Munich, Germany. It offers exact one- and two-sample tests and Hodges–Lehmann confidence intervals. Fisher's exact test is provided for the 2 × *c* contingency table. Only asymptotic tests are available for *r* × *c* contingency tables where $r > 2$.

References

1. Agresti A, Yang M (1987). An empirical investigation of some effects of sparseness in contingency tables. *Comm Stat* 5:9–21.
2. Read RC, Cressie NA (1988). *Goodness of Fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.
3. Fisher RA (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.
4. Agresti A (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.
5. Agresti A (1984). *Analysis of Ordinal Categorical Data*. New York: John Wiley.
6. Yates F (1984). Test of significance for 2 × 2 contingency tables. *J R Stat Soc A* 147: 426–463.
7. Cox DR, Hinkley DV (1974). *Theoretical Statistics*. London: Chapman and Hall.
8. Mehta CR, Patel NR (1983). A network algorithm for performing Fisher's exact test in *r* × *c* contingency tables. *J Am Stat Assoc* 78(382):427–434.
9. Freeman GH, Halton JH (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika* 38:141–149.
10. Mehta CR, Patel NR (1986). A hybrid algorithm for Fisher's exact test on unordered *r* × *c* contingency tables. *Comm Stat* 15(2):387–403.
11. Hollander M, Wolfe DA (1973). *Nonparametric Statistical Methods*. New York: John Wiley.
12. Agresti A, Mehta CR, Patel NR (1990). Exact inference for contingency tables with ordered categories. *J Am Stat Assoc* 85:410, 453–458.
13. Chernoff H, Savage IR (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Ann Math Stat* 29:972–994.
14. Gibbons JD (1985). *Nonparametric Statistical Inference*, 2nd edition. New York: Marcel Dekker.
15. Kalbfleish JD, Prentice RL (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons.
16. Hettmansperger TP (1984). *Statistical Inference Based on Ranks*. New York: John Wiley & Sons.
17. Breslow NE, Day NE (1980). The analysis of case-control studies. *IARC Scientific Publications No. 32*. Lyon, France: IARC.
18. Zelen M (1971). The analysis several 2 × 2 contingency tables. *Biometrika* 58(1):129–137.
19. Gart J (1970). Point and interval estimation of the common odds ratio in the combination of 2 × 2 tables with fixed marginals. *Biometrika* 57:471–475.
20. Cox DR, Snell EJ (1989). *The Analysis of Binary Data*, 2nd edition. New York: Chapman and Hall.
21. Mehta CR, Patel NR, Gray R (1985). On computing an exact confidence interval for the common odds ratio in several 2 × 2 contingency tables. *J Am Stat Assoc* 80(392):969–973.
22. Mantel N, Haenszel W (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 22:719–748.

23. Robins J, Breslow N, Greenland S (1986). Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* 42:311–323.
24. Mehta CR, Patel NR, Tsiatis AA (1984). Exact significance testing to establish treatment equivalence for ordered categorical data. *Biometrics* 40:819–825.
25. Mehta CR, Patel NR (1986). FEXACT: a Fortran subroutine for Fisher's exact test on unordered $r \times c$ contingency tables. *ACM Trans Math Software* 12(2):154–161.
26. Pagano M, Halvorsen K (1981). An algorithm for finding exact significance levels of $r \times c$ contingency tables. *J Am Stat Assoc* 76:931–934.
27. Pagano M, Tritchler D (1983). On obtaining permutation distributions in polynomial time. *J Am Stat Assoc* 78:435–441.
28. Daglivo J, Olivier D, Pagano M (1988). Methods for the analysis of contingency tables with large and small cell counts. *J Am Stat Assoc* 83:1006–1013.
29. Streitberg B, Rohmel R (1986). Exact distributions for permutation and rank tests. *Stat Software Newslett* 12:10–17.
30. Agresti A, Wackerly D (1977). Some exact conditional tests of independence for $r \times c$ cross-classification tables. *Psychometrika* 42:111–125.
31. Patefield WM (1981). An efficient method of generating $r \times c$ tables with given row and column totals. (Algorithm AS 159). *Appl Stat* 30:91–97.
32. Mehta CR, Patel NR,. Senchaudhuri P (1988). Importance sampling for estimating exact probabilities in permutational inference. *J Am Stat Assoc* 83(404):999–1005.
33. Gupta PC, Mehta FR, Pindborg J (1980). *Comm Dental Oral Epidemiol* 8:287–333.
34. Feynman RP (1988). *What Do You Care What Other People Think?* New York: W.W. Norton.
35. Gastwirth JL (1984). Combined tests of significance in EEO cases. *Indust Labor Rel Rev* 38(1).
36. Sprent P (1993). *Applied Nonparametric Statistical Methods*, 2nd edition. London: Chapman and Hall.
37. *EGRET User Manual* (1989). Statistics and Epidemiology Research Corporation, Seattle, WA.
38. *Epi Info Manual* (1989). Centers for Disease Control, Atlanta, GA.
39. *SAS/Stat Guide for Personal Computers* (1987). Version 6 edition. The SAS Institute, Cary, NC.
40. StatXact Version 3 (1993). *Software for Exact Nonparametric Inference.* Cytel Software Corporation, Cambridge, MA 02139.
41. Lehmann EL (1975). *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day.
42. Seigel S, Castellan NJ (1988). *Nonparametric Statistics for the Behavioral Sciences*, 2nd edition. New York: McGraw Hill.
43. LogXact (1993). *Software for Exact Logistic Regression.* Cytel Software Corporation, Cambridge, MA.
44. *Testimate Version 5.1* (1993). IDV, Datenanalyse und Versuchsplanung, Munich, Germany.

10. Stratified-adjusted versus unstratified assessment of sample size and power for analyses of proportions

John M. Lachin and Oliver M. Bautista

Introduction

In any scientific investigation, it is important to evaluate the adequacy of sample size with regard to one's ability to provide clear answers to the questions posed. In many cases, this assessment is based upon the power of a statistical test for the comparison of two groups with respect to the probability of some event or characteristic in two independent samples of subjects. In the simplest case, the proportions of subjects with some characteristic are compared between the two groups using a standard chi-square or Z-test for a 2 × 2 table. Various authors have described expressions for the approximate power of the large sample chi-square test, the most widely used being the expression based upon the large sample Z-test for two proportions of Halperin et al. [1]. This and other widely used procedures for the evaluation of sample size on the basis of power are reviewed by Lachin [2] and Donner [3], among others. This approach is based upon an unconditional or marginal assessment of the treatment group difference without consideration of other covariate effects.

For a single 2 × 2 table, the magnitude of the treatment effect can be expressed in terms of the odds ratio. To adjust for another qualitative covariate, or for the grouped categories of a quantitative covariate, the Mantel and Haenszel [4] procedure can be employed to obtain a stratified-adjusted estimator of the overall odds ratio and a stratified-adjusted test. A similar test, proposed by Cochran [5], employs the large-sample unconditional variance for the 2 × 2 table rather than the conditional variance as employed by Mantel–Haenszel. The Cochran–Mantel–Haenszel test was shown by Radhakrishna [6] to be asymptotically efficient against a sequence of local alternatives with a common odds ratio within each of the multiple 2 × 2 tables. Birch [7] described the noncentral distribution of the large-sample Mantel–Haenszel test from which the asymptotic power of the test could be assessed. More recently, Woolson, Bean, and Rojas [8] and Wittes and Wallenstein [9] described expressions for the power function of the

Peter F. Thall (ed.), RECENT ADVANCES IN CLINICAL TRIAL DESIGN AND ANALYSIS.
Copyright © 1995 Kluwer Academic Publishers, Boston. All rights reserved. ISBN 978-1-4613-5830-5.

Mantel–Haenszel test under different sampling designs by considering the expected values of the components of the test statistic under the null and alternative hypotheses. For the case of a single 2 × 2 table without stratification, these expressions reduce to the simple Halperin et al. [1] expression. These methods, therefore, allow for the assessment of sample size and power for a stratified-adjusted analysis of two proportions.

An alternate method for obtaining an adjusted assessment of treatment effect is to employ a regression model. A logistic regression model with binary indicator for treatment group provides the maximum likelihood estimate of the log odds ratio adjusting for the other covariate effects in the model. The power of the large sample Wald test for treatment effect can be approximated by the noncentral chi-square distribution with a noncentrality parameter based on the expected value of the Wald test [10,11]. Whittemore [12] provides an approximate solution for the power function of a logistic regression analysis in general, but this approach requires knowledge of the moment generating function of the covariates. Hsieh [13], Wilson and Gordon [14], and Self and Mauritsen [15] present generalizations of this method. All require either knowledge of or assumptions about the distribution of the covariate values that will form the design matrix in such models.

With these methods, it is possible to assess the power of a test for two proportions using stratification adjustment for other covariates or using a regression model adjustment for a collection of covariates. However, there has been little work to describe the conditions under which it is important, rather than superfluous, to consider a covariate adjustment for the assessment of sample size and power. Beach and Meier [16] and Canner [17], both using a model initially proposed by Canner [18], considered the effect of a covariance adjustment on a measure of treatment effect. Beach and Meier considered the difference between the adjusted and unadjusted Z-value for 2 × 2 × 2 tables, but only under the null hypothesis. Canner [17] considered the difference between the Z-statistics in a multiple regression model with a quantitative outcome measure. Both considered the case of a single baseline covariate, binary in the case of Beach and Meier [16], quantitative in the case of Canner [17]. Each showed that the difference between the adjusted and unadjusted Z-values is a function of the Z-values for the association between the covariate and the outcome, and of the Z-value for the association between the covariate and treatment group membership. These models, however, did not assess the impact of a treatment by covariate interaction on the response. In the case of multiple 2 × 2 tables, a treatment by covariate interaction is manifested by heterogeneity of the odds ratios among the 2 × 2 tables.

In this chapter, we describe extensive computations to assess the factors that affect the power of a test for two proportions with and without a stratification adjustment. These assessments will evaluate the effects of a covariate association with the response, covariate imbalance among treatment groups (covariate association with treatment group membership), and the extent of treatment by covariate interaction (heterogeneity of odds ratios over strata). We shall perform this evaluation using the large sample expression for the power of a Mantel–Haenszel test provided by Wittes and Wallenstein [9] for the case of two independent groups versus the unadjusted power from the marginal 2 × 2 table using the expression of Halperin et al. [1].

In the remainder of this chapter, we first describe the various measures of association and heterogeneity for multiple 2 × 2 tables. Then we describe the adjusted and unadjusted power functions for the test for two proportions. Next we describe the model under which the powers of these tests are compared. Then we describe the effects of different characteristics of a set of multiple 2 × 2 tables on the power of the adjusted and unadjusted tests. Finally, we present a discussion of the implications of the results obtained from these various computations.

Odds ratios for S 2 × 2 tables

Let n_{ijk} refer to the cell frequency for the jth response (success (+) or failure (−)) in the ith treatment group (experimental treatment (e) or control (c)) for subjects in the kth stratum ($k = 1, \ldots, S$). The total sample size is $N = \Sigma_{ijk} n_{ijk}$, and $E(n_{ijk}) = N\pi_{ijk}$, where π_{ijk} is the probability associated with the ijkth cell in the $2 \times 2 \times S$ table. Throughout we employ the '•' notation to designate summation over the corresponding index of the three-way table.

Conditionally, within the kth stratum, the 2 × 2 table is of the form

$$
\begin{array}{cc|c|c|c}
 & & \multicolumn{2}{c}{\text{group}} & \\
 & & \text{e} & \text{c} & \\
\cline{3-4}
\text{response} & + & n_{\text{e}+k} & n_{\text{c}+k} & n_{\bullet+k} \\
\cline{3-4}
 & - & n_{\text{e}-k} & n_{\text{c}-k} & n_{\bullet-k} \\
\cline{3-4}
 & & n_{\text{e}\bullet k} & n_{\text{c}\bullet k} & n_{\bullet\bullet k}
\end{array}
\tag{1}
$$

We assume that the total sample size, $n_{\bullet\bullet k}$, is known or is fixed by design, with stratum sample fraction $r_k = n_{\bullet\bullet k}/N$; $\Sigma_k r_k = 1.0$. Likewise, for the kth stratum, each treatment group sample size $n_{i\bullet k}$ has a corresponding fixed sample fraction $Q_{ik} = n_{i\bullet k}/n_{\bullet\bullet k}$ (i = e, c) where $Q_{ek} + Q_{ck} = 1$.

The *conditional association* between the treatment and response in stratum k is represented by the conditional odds ratio

$$\theta_{C_k} = \frac{\pi_{\text{e}+k}\pi_{\text{c}-k}}{\pi_{\text{e}-k}\pi_{\text{c}+k}}. \tag{2}$$

For the case of only two strata ($S = 2$), the *heterogeneity* of treatment group-response association among strata is represented by

$$\psi = \theta_{C_1}/\theta_{C_2}. \tag{3}$$

Three different 2-way marginal tables can then be constructed. One describes the *pooled* or *unadjusted association* between treatment and response

		group			
		e	c		
response	+	$n_{e+\bullet}$	$n_{c+\bullet}$	$n_{\bullet+\bullet}$	(4)
	−	$n_{e-\bullet}$	$n_{c-\bullet}$	$n_{\bullet-\bullet}$	
		$n_{e\bullet\bullet}$	$n_{c\bullet\bullet}$	N	

with cell expectations $\{N\pi_{ij\bullet}\}$. Here the total sample sizes $n_{i\bullet\bullet}$ are assumed known or fixed by design with corresponding sample fractions $Q_i = n_{i\bullet\bullet}/N$ (i = e, c). Also, $E(n_{\bullet+\bullet}) = N\pi_{\bullet+\bullet}$, where $\pi_{\bullet+\bullet}$ is the overall probability or prevalence of a positive response in the population. The unadjusted odds ratio is provided by

$$\theta_U = \frac{\pi_{e+\bullet}\pi_{c-\bullet}}{\pi_{e-\bullet}\pi_{c+\bullet}}. \tag{5}$$

The $S \times 2$ marginal table of stratum-by-group describes the imbalance between treatment groups among strata:

		group			
		e	c		
	1	$n_{e\bullet 1}$	$n_{c\bullet 1}$	$n_{\bullet\bullet 1}$	
stratum	2	$n_{e\bullet 2}$	$n_{c\bullet 2}$	$n_{\bullet\bullet 2}$	(6)
	⋮	⋮	⋮	⋮	
	S	$n_{e\bullet S}$	$n_{c\bullet S}$	$n_{\bullet\bullet S}$	
		$n_{e\bullet\bullet}$	$n_{c\bullet\bullet}$	N	

where $n_{\bullet\bullet k} = r_k N$ ($k = 1, \ldots, S$) and $n_{i\bullet\bullet} = Q_i N$ (i = e, c). For only two strata, the group-by-stratum *imbalance* can be measured by the odds ratio

$$\theta_I = \frac{\pi_{e\bullet 1}\pi_{c\bullet 2}}{\pi_{e\bullet 2}\pi_{c\bullet 1}}. \tag{7}$$

For S strata, the association can be described by a vector of S-1 odds ratios of the first stratum versus each of the remaining strata ($k = 2, \ldots, S$).

Likewise, the $S \times 2$ marginal table of stratum-by-response describes the association between stratum and response:

$$\begin{array}{ll|c|c|l} & & \multicolumn{2}{c}{\text{response}} & \\ & & + & - & \\ \cline{3-4} & 1 & n_{\bullet+1} & n_{\bullet-1} & n_{\bullet\bullet1} \\ \cline{3-4} \text{stratum} & 2 & n_{\bullet+2} & n_{\bullet-2} & n_{\bullet\bullet2} \\ \cline{3-4} & \vdots & \vdots & \vdots & \vdots \\ \cline{3-4} & S & n_{\bullet+S} & n_{\bullet-S} & n_{\bullet\bullet S} \\ \cline{3-4} & & n_{\bullet+\bullet} & n_{\bullet-\bullet} & N \end{array} \tag{8}$$

where $E(n_{\bullet+\bullet}) = N\pi_{\bullet+\bullet}$. For only two strata, the *stratum-response association* is measured by

$$\theta_A = \frac{\pi_{\bullet+1}\pi_{\bullet-2}}{\pi_{\bullet-1}\pi_{\bullet+2}}. \tag{9}$$

Each of the odds ratios can be estimated by the corresponding cross-product ratio of the cell frequencies. However, if we assume that the conditional odds ratios have a common expectation such that $\theta_{C_1} = \theta_{C_2} = \cdots = \theta_{C_s} = \bar{\theta}$, then the maximum likelihood estimate of $\ln(\bar{\theta})$ can be obtained from a logistic regression model adjusting for stratum. This requires an iterative solution, even for the 2 × 2 × 2 table. In general, under the stratified model, this odds ratio will differ from the unadjusted (pooled) odds ratio, even when where is a common odds ratio $\bar{\theta}$ within strata.

Sample size and power: adjusted vs. unadjusted

The Mantel and Haenszel [4] and Cochran [5] test statistics for a stratified-adjusted analysis of S 2 × 2 tables each employ a weighted average of the differences in the sample proportions in each 2 × 2 table, $d_k = (n_{e+k}/n_{e\bullet k}) - (n_{c+k}/n_{c\bullet k})$, rather than the sample odds or log odds ratios. Similarly, Wittes and Wallenstein [9] described an approximation to the power of the Mantel–Haenszel test that is also based on a weighted sum of the differences within strata:

$$\delta_k = \frac{\pi_{e+k}}{\pi_{e\bullet k}} - \frac{\pi_{c+k}}{\pi_{c\bullet k}} = \frac{\pi_{e+k}}{r_k Q_{ek}} - \frac{\pi_{c+k}}{r_k Q_{ck}} \tag{10}$$

An equivalent expression for the power of this test was also presented by Woolson, Bean, and Rojas [8] in the setting of a stratified case–control study.

In general, the equation relating sample size and power for a normally distributed test statistic with expectation μ is of the form

$$\sqrt{N}\,|\Delta_\mu| = Z_{1-\alpha}\Sigma_0 + Z_{1-\beta}\Sigma_1 \tag{11}$$

(cf. Lachin [2]), where Δ_μ is the location difference under the alternative, and the variance of the test statistic is Σ_0^2/N under the null hypothesis, Σ_1^2/N under the alternative.

The *unadjusted* test for two proportions is based on the treatment-by-response marginal table (4). For this table, let $p_{e\bullet} = (\pi_{e+\bullet}/Q_e)$ and $p_{c\bullet} = (\pi_{c+\bullet}/Q_c)$. Then the sample size and power of the test are provided by equation (11) with

$$\begin{aligned}
\Delta_\mu &= p_{e\bullet} - p_{c\bullet}\\
\Sigma_0^2 &= \frac{\bar{p}_\bullet(1-\bar{p}_\bullet)}{Q_c Q_e},\ \bar{p}_\bullet = \pi_{e+\bullet} + \pi_{c+\bullet}\\
\Sigma_1^2 &= \frac{p_{e\bullet}(1-p_{e\bullet})}{Q_e} + \frac{p_{c\bullet}(1-p_{c\bullet})}{Q_c}
\end{aligned} \tag{12}$$

which yields the expression of Halperin et al. [1].

For the stratified-adjusted test, within the kth table let $p_{ik} = \pi_{i+k}/(Q_{ik}r_k)$ (i = e, c). Then the sample size and power of the stratified-adjusted Mantel–Haenszel test are provided by equation (11) with

$$\begin{aligned}
\Delta_\mu &= \Sigma_k r_k Q_{ek} Q_{ck}(p_{ek} - p_{ck})\\
\Sigma_0^2 &= \Sigma_k r_k Q_{ek} Q_{ck} \bar{p}_k(1-\bar{p}_k),\ \bar{p}_k = Q_e p_{ek} + Q_c p_{ck}\\
\Sigma_0^2 &= \Sigma_k r_k Q_{ek} Q_{ck}[Q_{ck} p_{ek}(1-p_{ek}) + Q_{ek} p_{ck}(1-p_{ck})].
\end{aligned} \tag{13}$$

For a single 2 × 2 table, equations (13) reduce to equations (12).

For the calculation of sample size a priori, one specifies the stratum sample fractions $\{r_k\}$ and the stratum experimental treatment group fractions $\{Q_{ek}\}$ (or control $\{Q_{ck}\}$) and the cell probabilities $\{\pi_{ijk}\}$ subject to these marginal constraints. One then solves for N. Alternatively, and more accurately, the computations could be performed in terms of the cell frequencies $\{n_{ijk}\}$. Greenland [19] used this approach, assuming that the relative risk π_{e+k}/π_{c+k} is the same for each k. Minor discrepancies may arise because for a given N, the expected frequency $N\pi_{ijk}$ generally is not integer valued. One can generate the table of expected frequencies under some model, and then assess sample size or power using the corresponding sample proportions $\hat{\pi}_{ijk} = n_{ijk}/N$.

Power of the unadjusted versus stratified-adjusted tests

In some instances, the power of the unadjusted test equals that of the stratified-adjusted test, while in other cases the two can be vastly different. We now describe a model for multiple 2 × 2 tables under which we will assess the factors that affect the power of the adjusted versus the unadjusted test. For simplicity, we consider only the case of two strata (a 2 × 2 × 2 table).

For a given N, a 2 × 2 × 2 table can be parameterized as a function of the following quantities:

a. the group sample size $n_{e\bullet\bullet} = Q_e N$
b. the overall prevalence of a positive response $n_{\bullet+\bullet} = \pi_{\bullet+\bullet} N$
c. the unadjusted odds ratio θ_U in equation (5) in terms of the $\{\pi_{ij\bullet}\}$ or the corresponding $\{n_{ij\bullet}\}$
d. the sample size of the first stratum $n_{\bullet\bullet 1} = r_1 N$(and thus $n_{\bullet\bullet 2} = N - n_{\bullet\bullet 1}$).
e. the odds ratio for the imbalance θ_I in equation (7)
f. the odds ratio for stratum-response association θ_A in equation (9)
g. the heterogeneity ratio of within-stratum odds ratio ψ in equation (3).

From these seven quantities, the complete 2 × 2 × 2 table of probabilities or expected frequencies can be generated (see the appendix). Briefly, from (a), (b), and (c), the pooled table entries (4) are obtained. From (d) and (e), the stratum-by-group table (6) is generated. From (f) and the other known margins, the stratum-by-response table (8) is generated. From (g), the n_{e1+} cell is obtained, from which the two separate 2 × 2 tables are also generated, with corresponding odds ratios $\{\theta_{C_k}\}$.

From the pooled table entries given in table (4), the power of the unadjusted test can be computed from equations (11) and (12). Let this power be denoted by $\beta_U^c = 1 - \beta_U$, where β_U is the unadjusted type II error. Similarly, from the stratum 1 and stratum 2 tables given in (1), the power of the Mantel–Haenszel test adjusted for strata can be computed using Wittes and Wallenstein's procedure from equations (11) and (13). Let this stratified-adjusted power be denoted by $\beta_A^c = 1 - \beta_A$, where β_A is the adjusted type II error. The following investigation evaluates the influence of the above seven quantities on the difference in power of the unadjusted versus the stratified-adjusted analyses. In all cases, the probability of type I error is fixed at 0.05 (two-sided).

An initial investigation was conducted to determine the effect, if any, of the degree of stratum-by-group imbalance θ_I, stratum-by-response association θ_A, and heterogeneity of the treatment effect over strata ψ, on the difference between the unadjusted and stratified-adjusted powers. To investigate a reasonably wide range of possible situations, while controlling the overall size of the initial investigation, we considered the following parameter values:

N	= 100, 200, 400, 800	θ_U	= 1, 1.25, 1.5, 1.75, 2, 3, 4, 5
Qe	= 0.5	θ_I	= 0.2, 1, 5
$\pi_{\bullet+\bullet}$	= 0.1, 0.2, 0.3, 0.4, 0.5	θ_A	= 0.2, 1, 5
r_1	= 0.25, 0.5, 0.75	ψ	= 0.2, 1, 5

For a given combination of the values for N, Qe, $\pi_{\bullet+\bullet}$, r_1, θ_U, θ_I, θ_A, and ψ, the pooled, stratum 1 and stratum 2 tables were constructed as outlined in the appendix. From these tables, the unadjusted and stratified-adjusted powers β_U^c and β_A^c were obtained. The following working definition of a "difference" between β_U^c and β_A^c was adopted:

Definition 1. Let $\Delta_{\beta^c} = \beta_A^c - \beta_U^c$. For a fixed N, Qe, $\pi_{\bullet+\bullet}$, r_1, and ψ, a 'difference' between β_A^c and β_U^c is defined to exist if for at least one θ_U in $\{1,2,3,4,5\}$, the corresponding $|\Delta_{\beta^c}|$ is greater than 0.1.

This approach attempts to capture those situations in which there is an algebraic difference of at least 10% between the unadjusted and stratified-adjusted powers. Here, the focus is on whether a moderate difference in power exists, not on its magnitude.

For each of the possible combinations of N, Qe, $\pi_{\bullet+\bullet}$, and r_1 in the range given above, table 1 gives a summary of the values of θ_I, θ_A, and ψ where a difference does and does not exist between the unadjusted and stratified-adjusted powers. Table 1 suggests that a certain degree of stratum-by-group imbalance and stratum-by-response association are required in order to have a difference of more than 10% between the unadjusted and stratified-

Table 1. Result of initial investigation

θ_I	θ_A	ψ	Power difference?
0.2	0.2	0.2	Yes
		1.0	Yes
		5.0	Yes
	1.0	0.2	No
		1.0	No
		5.0	No
	5.0	0.2	Yes
		1.0	Yes
		5.0	Yes
1.0	0.2	0.2	No
		1.0	No
		5.0	No
	1.0	0.2	No
		1.0	No
		5.0	No
	5.0	0.2	No
		1.0	No
		5.0	No
5.0	0.2	0.2	Yes
		1.0	Yes
		5.0	Yes
	1.0	0.2	No
		1.0	No
		5.0	No
	5.0	0.2	Yes
		1.0	Yes
		5.0	Yes

adjusted powers. Table 1 also shows that a difference exits between the adjusted and unadjusted powers for values of ψ equal to 0.2, 1, and 5 only when there is a certain degree of imbalance and association. In fact, when there is no imbalance ($Q_{ek} = Q_e$ for all strata), for the stratified and unstratified assessment of power in equations (12) and (13), it is readily shown that the location shifts Δ_μ are equal and that there is only a slight difference in the variance components Σ_1. Thus, the presence or absence of heterogeneity between strata alone does not materially affect whether there is a difference between the adjusted and unadjusted powers.

Maximum likelihood estimate of the common stratum odds ratio θ and the unadjusted odds ratio θ_U

As mentioned in an earlier section, when the conditional odds ratios θ_{C_1} and θ_{C_2} have a common expectation, say $\bar{\theta}$, in general, this common odds ratio will differ from the unadjusted odds ratio θ_U. In the case of a substantial difference, in most cases one would infer that the unadjusted odds ratio is biased and that the adjusted odds ratio provides an unbiased assessment of the treatment group effect. For a given set of 2 × 2 tables, $\bar{\theta}$ can be obtained from a logistic regression model as follows. Let T be an indicator variable for treatment group ($T = +1$ if e, 0 if c) and let X be another indicator for stratum (0 if stratum 1, 1 if stratum 2). Then in a logistic model of the form $\ln[p/(1 + p)] = \alpha + \beta T + \gamma X$, $\exp(\beta)$ provides the maximum likelihood estimate of $\bar{\theta}$.

The relationship between the common or adjusted odds ratio $\bar{\theta}$ and the unadjusted odds ratio θ_U is now investigated for the cases when $N = 800$, $Q\text{e} = 0.5$, $\pi_{\bullet+\bullet} = 0.5$, $r_1 = 0.5$, $\psi = 1$, $\theta_I = \{1/5,1/3,1,3,5\}$, and $\theta_A = \{1/5, 1/3,1,3,5\}$. Since $\psi = 1$ by design, $\bar{\theta}$ is the maximum likelihood estimate of the common odds ratio of stratum 1 and stratum 2.

Figure 1 shows the plots of $\bar{\theta}$ versus θ_U, with the solid line being the line of equality. When no covariate imbalance or association is present, as shown by the '$(\theta_I, \theta_A) = (5, 1)$ or $(1, 5)$' curve, the common odds ratio $\bar{\theta}$ is always greater than the unadjusted odds ratio θ_U. However, when there is substantial covariate imbalance and association, the common odds ratio $\bar{\theta}$ is always less than the unadjusted odds ratio θ_U. In general, $\bar{\theta}$ increases with θ_U — and furthermore, the difference between $\bar{\theta}$ and θ_U also increases. This indicates that the bias in the unadjusted odds ratio increases linearly as θ_U increases. However, as will be shown in the next section, this does not translate into a monotonic difference in the stratified-adjusted versus unadjusted powers as θ_U increases.

Note that the direction of the covariate imbalance or association does not change the relationship between the adjusted and unadjusted odds ratio. For instance, the same $\bar{\theta} \times \theta_U$ curve is obtained for both $(\theta_I,\theta_A) = (3,3)$ and $(\theta_I,\theta_A) = (1/3,1/3)$.

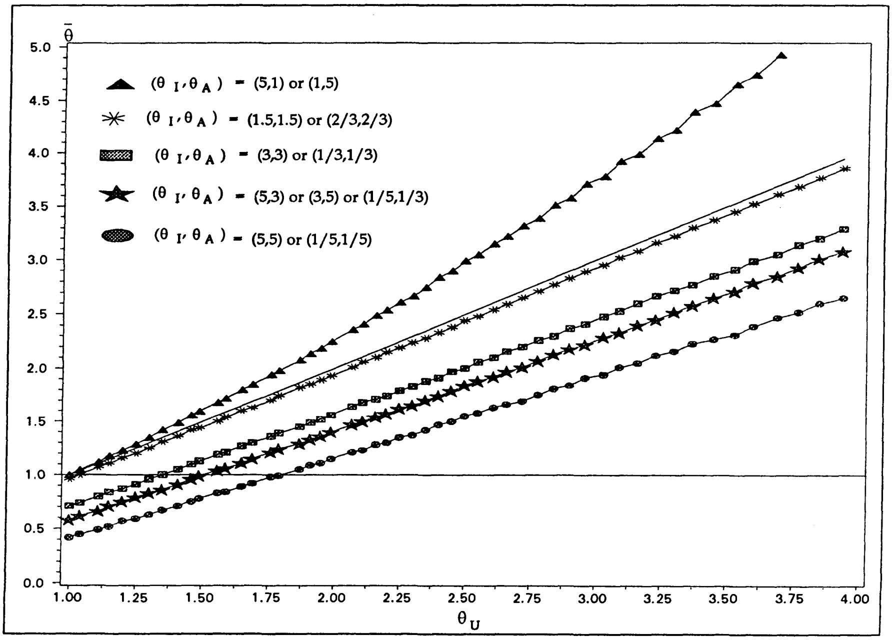


Figure 1. Plots of the adjusted odds ratio $\bar{\theta}$ versus the unadjusted odds ratio θ_U for $N = 800$, $Qe = 0.5$, $\pi_{\bullet+\bullet} = 0.5$, $r_1 = 0.5$, and $\psi = 1$.

Effect of the unadjusted odds ratio θ_U on the power difference

We now consider the effect of the value of the unadjusted odds ratio θ_U, and indirectly that of the adjusted odds ratio, on the difference between the unadjusted and stratified-adjusted powers. In general, for a fixed value of the other parameters, the difference between the adjusted and unadjusted powers varies as a function of the unadjusted odds ratio θ_U. Figure 2 shows the plot of the power difference $\Delta_{\beta^c} = B^c_A - \beta^c_U$ as a function of the unadjusted odds ratio θ_U, θ_I, and θ_A when $N = 800$, $Qe = 0.5$, $\pi_{\bullet+\bullet} = 0.5$, $r_1 = 0.5$, and $\psi = 1$. As in the plot of $\bar{\theta}$ versus θ_U, the direction of the imbalance or association does not affect the power difference Δ_{β^c} as shown, for example, by the same curve for $(\theta_I, \theta_A) = (5,5)$ and $(\theta_I, \theta_A) = (1/5,1/5)$. Also, virtually identical values are obtained when there is heterogeneity among strata ($\psi = 5$).

The largest absolute power difference generally occurs when $\theta_U \in [1,2]$ and becomes negligible for higher values of θ_U. When $(\theta_I, \theta_A) = (5,5)$, the largest absolute power difference occurs when $\theta_U = 1$, while when $(\theta_I, \theta_A) = (3,3)$, the largest absolute difference occurs when $\theta_U = 1.5$.

A positive difference indicates that the stratified-adjusted power is greater

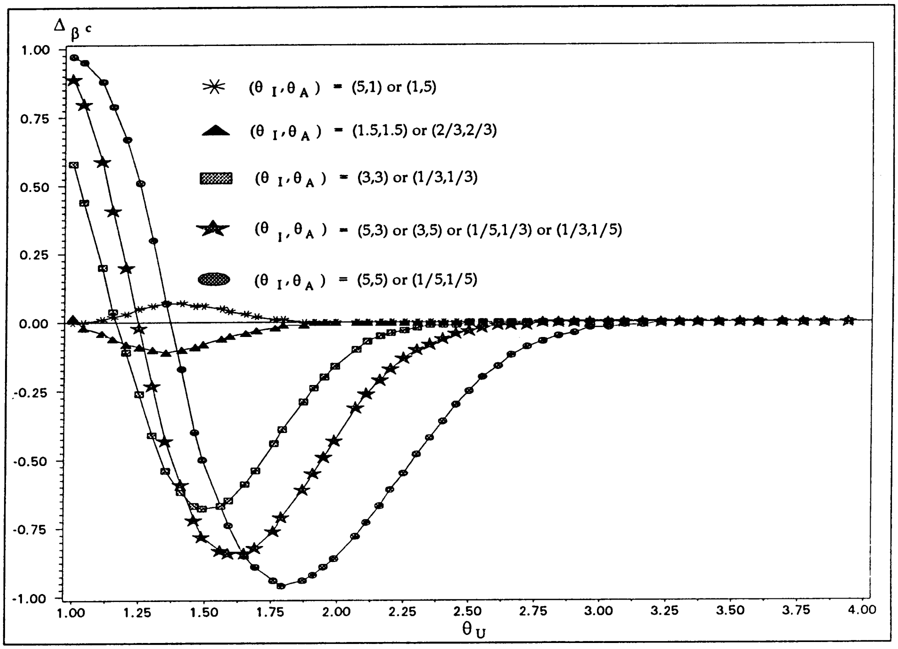


Figure 2. Plots of the power difference $\Delta_{\beta c}$ versus the unadjusted odds ratio θ_U for $N = 800$, $Qe = 0.5$, $\pi_{\bullet + \bullet} = 0.5$, $r_1 = 0.5$, and $\psi = 1$.

than the unadjusted power. Figure 2 indicates that for $(\theta_I, \theta_A) = (5,5)$ — and for $(\theta_I, \theta_A) = (1/5,1/5)$ — a positive difference occurs (approximately) when $\theta_U \in [1.0,1.35]$, while a negative difference occurs (approximately) when $\theta_U \in [1.4,3.0]$. When $\theta_U > 3.0$, the difference is already negligible. Thus, for the case when $N = 800$, $Qe = 0.5$, $\pi_{\bullet + \bullet} = 0.5$, $r_1 = 0.5$, $\psi = 1$, and $\theta_I = \theta_A = 5$, when the unadjusted odds ratio is less than 1.4, the unadjusted approach will tend to underestimate power and overestimate the required sample size relative to the more accurate stratified approach. For values above 1.4, the opposite occurs. Here the unstratified approach provides greater power due to the positive bias of the unadjusted estimate, i.e., power is overestimated and sample size underestimated, this difference reaching a maximum at about $\theta_U = 1.79$.

It is instructive to examine some specific cases when $\theta_I = \theta_A = 5$. When the unadjusted odds ratio is 1, due to the substantial covariate imbalance and association, the stratified-adjusted odds ratio is substantially different from 1, and thus substantial power is lost by not adjusting for the stratum effect. For example, when $N = 800$, $Qe = 0.5$, $\pi_{\bullet + \bullet} = 0.5$, $r_1 = 0.5$, $\theta_I = \theta_A = 5$, $\psi = 1$, and $\theta_U = 1$, the corresponding stratum 1, stratum 2, and pooled 2 × 2 tables are as follows:

S1	e	c	Total
+	176	100	276
−	100	24	124
Total	276	124	400

$\theta_{C_1} = 0.4224 =$ stratum 1 odds ratio

S2	e	c	Total
+	24	100	124
−	100	176	276
Total	124	276	400

$\theta_{C_2} = 0.4224 =$ stratum 2 odds ratio

	e	c	Total
+	200	200	400
−	200	200	400
Total	400	400	800

$\theta_U = 1.0 =$ pooled table odds ratio

Both stratum 1 and stratum 2 have an odds ratio of 0.4224. For the Mantel–Haenszel test, this yields an adjusted location difference Δ_{μ_A} of 0.03636 in equations (13). However, when the two tables are combined, the resulting pooled table produces an odds ratio of 1 and an unadjusted location difference Δ_{μ_U} in equations (12) of zero. Therefore, these discrepancies in the odds ratios and location differences for the stratum-specific tables are reflected in the unadjusted and stratified-adjusted powers.

On the other extreme, the lowest point of the '$(\theta_I,\theta_A) = (5,5)$' curve in figure 2 occurs when $\theta_U \approx 1.79$. The corresponding stratum 1, stratum 2, and pooled 2 × 2 tables are as follows:

S1	e	c	Total
+	191	85	276
−	85	39	124
Total	276	124	400

$\theta_{C_1} =$ stratum 1 odds ratio ≈ 1.0

S2	e	c	Total
+	38	86	124
−	86	190	276
Total	124	276	400

$\theta_{C_2} =$ stratum 2 odds ratio ≈ 1.0

	e	c	Total
+	229	171	400
−	171	229	400
Total	400	400	800

θ_U = pooled table odds ratio $\approx$ 1.79

In this case, both stratum 1 and stratum 2 have an odds ratio of 1, whereas the pooled table has an odds ratio of 1.79. In this case, due to the extreme imbalance of treatment group and stratum ($\theta_I = 5$), most would agree that the true odds ratio is 1.0, not 1.79. Thus, the unadjusted analysis is biased away from the null. As a result, the unadjusted analysis severely overestimates power relative to an unbiased stratified-adjusted analysis.

Even though the unadjusted odds ratio is nearly constantly positively biased when θ_I and θ_A are not equal to 1, the difference in the adjusted and unadjusted powers becomes negligible beyond some point. This suggests some difference either in the location shift parameters or variances from which the power functions are estimated.

Figure 3 shows the plots of the difference $\delta_{\Delta_\mu} = |\Delta_\mu|_A - |\Delta_\mu|_U$ as a function of the unadjusted odds ratio θ_U for $N = 800$, $Qe = 0.5$, $\pi_{\bullet+\bullet} =$

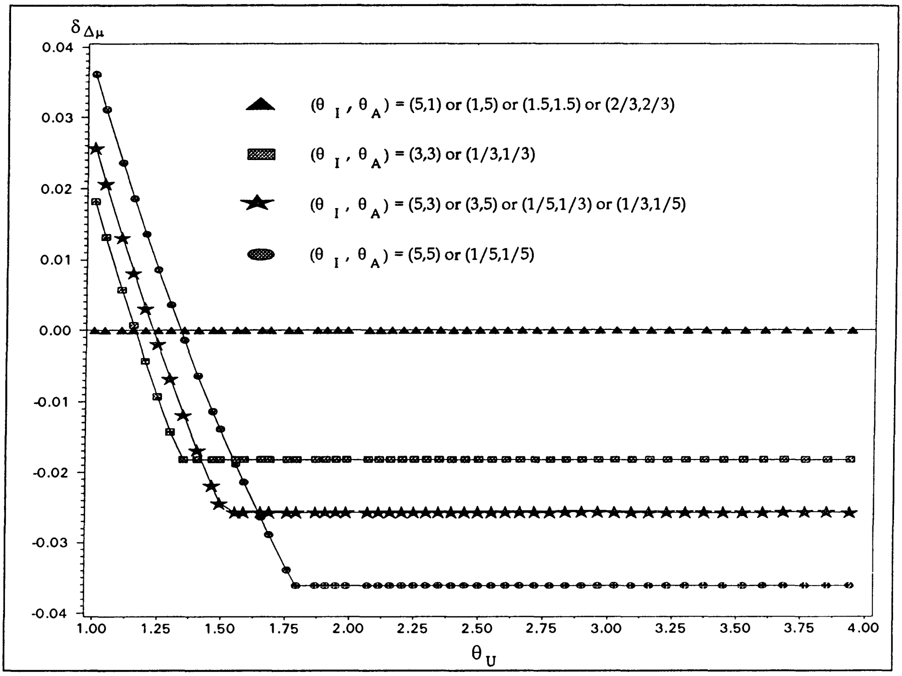


Figure 3. Plots of the difference δ_{Δ_μ} vs. the unadjusted odds ratio θ_U for $N = 800$, $Qe = 0.5$, $\pi_{\bullet+\bullet} = 0.5$, $r_1 = 0.5$, and $\psi = 1$.

0.5, $r_1 = 0.5$, $\psi = 1$, and a range of values for (θ_I,θ_A). It shows the range of θ_U where the difference between $|\Delta_\mu|_A$ and $|\Delta_\mu|_U$ contributes a large part to the difference between the adjusted and unadjusted powers. For the case when $(\theta_I,\theta_A) = (5,5)$, the positive difference when $\theta_U \in [1.0,1.3]$ corresponds to the cases in which the adjusted power is larger than the unadjusted power. Figure 1 shows that the adjusted odds ratio $\bar{\theta}$ is actually less than the null value 1.0 in this range of θ_U. This reflects the case in which the unadjusted analysis is biased towards the null. The negative difference when $\theta_U \in [1.30,1.79]$ corresponds to those cases in which the unadjusted power is larger than the stratified-adjusted power. Again, Figure 1 shows that in this range, the adjusted odds ratio $\bar{\theta}$ approaches 1.0 while the unadjusted odds ratio θ_U is biased away from 1.0.

The difference δ_{Δ_μ} for the case when $(\theta_I,\theta_A) = (5,5)$ is already constant when $\theta_U > 1.79$. In figure 1, this is the point at which both the adjusted odds ratio $\bar{\theta}$ and the unadjusted odds ratio θ_U are greater than 1.0, away from the null. Thus, for these cases, the difference between the adjusted and unadjusted location shifts is constant.

The point θ_U where the Δ_{β^c} curve in figure 2 reaches its lowest point differs for different degrees of covariate imbalance and association. Note that the shapes of the Δ_{β^c} curve in figure 2 and of the δ_{Δ_μ} curve in figure 3 shift to the left and taper off when θ_I or θ_A or both decrease from 5 to 3. The point θ_U, where the largest absolute power difference occurs, also differs for different degrees of covariate imbalance and association. In cases where a substantial difference is observed between the adjusted and unadjusted power, a substantial covariate imbalance and association is also present. When no substantial covariate imbalance or association is present, there is also no substantial difference in the adjusted and unadjusted powers.

We now examine the influence of the variance components. Let $\Sigma^2_{1_U}$ denote the unadjusted variance component Σ^2_1 in equations (12), and let $\Sigma^2_{1_A}$ denote the adjusted variance component Σ^2_1 in equations (13). Figure 4 shows plots of the difference $\Delta_\Sigma = \Sigma_{1_A} - \Sigma_{1_U}$ as a function of θ_U when $N = 800$, $Qe = 0.5$, $\pi_{\bullet+\bullet} = 0.5$, $r_1 = 0.5$, and $\psi = 1$. In the particular cases investigated, although Δ_Σ is strictly increasing as a function of θ_U when there is substantial covariate imbalance and association, the increasing part of the curves, say in the range [1.0,1.79] when $(\theta_I,\theta_A) = (5,5)$, is not reflected in figure 2 because this part is dominated by the decreasing value of δ_{Δ_μ}. For $\theta_U > 1.79$, when $(\theta_I,\theta_A) = (5,5)$, δ_{Δ_μ} is constant, so the increasing value of Δ_Σ is now reflected in the plot of Δ_{β^c} in figure 2.

The difference in the variance component Σ_0 in (14) for the unadjusted and stratified-adjusted analysis does not contribute to the Δ_{β^c} curve in figure 2. If we denote by $\Sigma^2_{0_U}$ and $\Sigma^2_{0_A}$ the unadjusted and adjusted variance components Σ^2_0 in equations (12) and (13), respectively, the difference $\Sigma_{0_A} - \Sigma_{0_U}$ is constant as a function of θ_U.

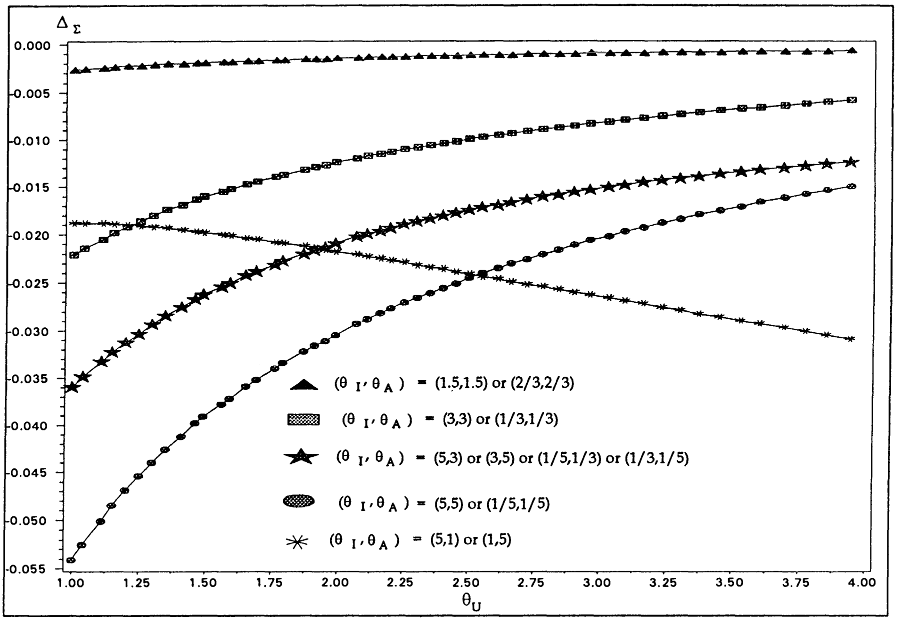


Figure 4. Plots of the difference Δ_Σ vs. the unadjusted odds ratio θ_U for $N = 800$, $Qe = 0.5$, $\pi_{\bullet+\bullet} = 0.5$, $r_1 = 0.5$, and $\psi = 1$.

Effect of the overall prevalence rate $\pi_{\bullet+\bullet}$ and the stratum fraction r_1

When there is substantial covariate imbalance and association, it can be seen in figure 2 that there is a point θ_U^* beyond which the difference between the unadjusted and stratified-adjusted powers approaches 0. For example, for the $(\theta_I,\theta_A) = (5,5)$ case shown in figure 2, $\theta_U^* \approx 3.20$. Let the point θ_U^* be referred to as the 'asymptote.' Some of the variables that influence the magnitude of this asymptote are explored in this section. As before, we fix N at 800 and Qe at 0.5.

The investigations conducted in the previous section regarding the effect of θ_I and θ_A on the Δ_{β^c} curve showed that the deflections in the curve shift to the right as θ_I or θ_A or both increase, and to the left as they decrease. In order to investigate the effects of the other factors on the asymptote θ_U^*, θ_I and θ_A are both fixed at 5.

Investigations in the previous sections showed that ψ does not materially affect the difference between the unadjusted and stratified-adjusted powers. This translates to a negligible effect on the Δ_{β^c} curve in figure 2. Since ψ has no apparent 'shifting' effect on the Δ_{β^c} curve, for fixed values of N, Qe, $\pi_{\bullet+\bullet}$, and r_1, with $\theta_I = \theta_A = 5$, changing ψ will not materially affect the value of the asymptote θ_U^*. For this reason, only the case in which $\psi = 5$ is considered for this particular investigation.

To investigate the effect of $\pi_{\bullet+\bullet}$ and r_1 on θ_U^*, $\pi_{\bullet+\bullet}$ is varied in the interval [0.2,0.9] in increments of 0.1, while r_1 is varied in the interval [0.1,0.9] in increments of 0.1. For $N = 800$, $Qe = 0.5$, $\theta_I = \theta_A = \psi = 5$, and for a fixed $\pi_{\bullet+\bullet}$ and r_1, the following working definition of 'asymptote' is adopted.

Definition 2. Let θ_U vary between 1.0 to 4.0 in increments of 0.05. For each θ_U, compute the corresponding $\Delta_{\beta^c} = \beta_A^c - \beta_U^c$. Define the set S as the collection of all θ_U such that the $|\Delta_{\beta^c}|$ associated with all such θ_U is greater than 0.005. The set S consists of all points θ_U where the unadjusted and stratified-adjusted power differ algebraically by (approximately) at least 1%. The *asymptote* is then defined as $\theta_U^* = \max_{\theta_U \in S}\{\theta_U\}$.

Figure 5 shows the surface plot of θ_U^* corresponding to the points $(r_1,\pi_{\bullet+\bullet})$, for $N = 800$, $Qe = 0.5$, $\theta_I = \theta_A = \psi = 5$. The surface plot is an asymmetric saddle. As r_1 varies, the asymptote θ_U^* is maximum at $r_1 \approx 0.50$, while as $\pi_{\bullet+\bullet}$ varies, it is minimum at $\pi_{\bullet+\bullet} \approx 0.50$. Thus the saddle point occurs in the neighborhood of the point (0.5,0.5). Note that θ_U^* ranges from a low of 2.19 to a high of 4.98.

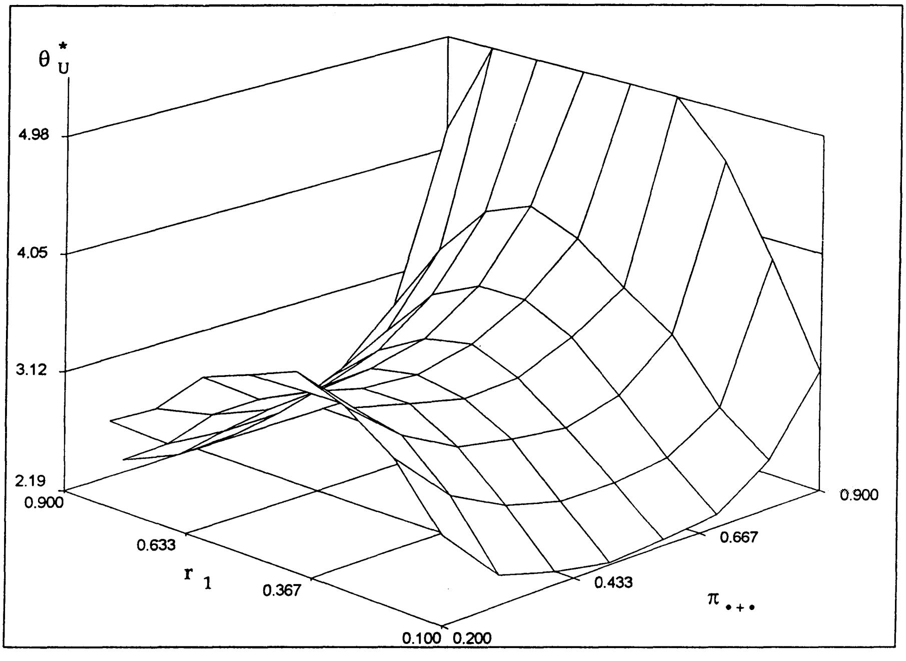


Figure 5. Surface plot of the asymptote θ_U^* corresponding to points $(r_1,\pi_{\bullet+\bullet})$ for $N = 800$, $Qe = 0.5$, and $\theta_I = \theta_A = \psi = 5$.

From figure 5, one can infer the effect of the stratum fraction r_1 and the overall prevalence rate $\pi_{\bullet+\bullet}$ on the difference Δ_{β^c} between the stratified-adjusted and unadjusted powers. An increase in θ_U^* would translate to a right-shift of the Δ_{β^c} curve in figure 2, while a decrease in θ_U^* would translate to a left-shift of the Δ_{β^c} curve. For a fixed stratum fraction r_1, a maximal left-shift occurs when $\pi_{\bullet+\bullet} = 0.50$. The opposite effect occurs when $\pi_{\bullet+\bullet}$ is fixed and r_1 is varied. For a fixed overall prevalence rate $\pi_{\bullet+\bullet}$, a maximal right-shift occurs at $r_1 = 0.50$. The principal effect of these shifts in the curve in figure 2 is to affect extrema with respect to θ_U with little effect on the extent of the amplitudes of the curve.

Discussion

These computations showed that a substantial difference in power between the unadjusted versus a stratified-adjusted analysis arises only when there is a certain degree of stratum-by-group imbalance and stratum-by-response association (table 1). A stratified-adjusted analysis will not provide a substantially different power when a covariate imbalance exists, but there is little or no covariate association with response. Similarly, there is no difference in power when covariate association exists, but there is no covariate imbalance. The presence or absence of heterogeneity between the stratum odds ratios does not materially affect the difference between the unadjusted and stratified-adjusted powers.

When both a covariate association and a covariate imbalance exists, however, the unadjusted odds ratio is positively biased relative to the stratum-adjusted odds ratio (figure 1). This bias is nearly constant over the range of the values of the unadjusted odds ratio and increases as a function of the extent of imbalance and association. Both imbalance and association must be present to some degree for there to be a substantial bias for relatively small values of θ_U.

This bias in the unadjusted odds ratio translates into a difference in the location shift parameters Δ_μ as θ_U increases, but only to a point beyond which no further change occurs (figure 3). Again, the shape of these curves depends on the extent of imbalance and association, there being no difference in these location shift parameters when there is no covariate association or no covariate imbalance (either θ_I or θ_A equal to 1.0). On the other hand, the expressions for the variance under the alternative hypothesis show an increasing difference as a function of θ_U when there is covariate association and imbalance, and a slightly declining difference when there is either no association or no imbalance (figure 4).

These two effects yield a sinusoidal-like curve for the difference between the power of the adjusted versus the unadjusted test as a function of θ_U (figure 2). Since an unadjusted odds ratio equal to 1.0 (the null, unadjusted) is positively biased, the stratified-adjusted test will yield greater power because the adjusted odds ratio in fact differs from the null in the opposite direction (figure 1). As θ_U increases, the adjusted odds ratio $\bar{\theta}$ approaches the null and then exceeds it, but θ_A is less than θ_U by a nearly constant amount. Thus, at some point in the neighborhood of $\theta_U \in [1.25,1.50]$ for the values shown, the unadjusted analysis provides greater power. However, the difference in the location shifts becomes a constant for a value of $\theta_U \in [1.5, 1.8]$ (figure 3) and thereafter the difference in powers is a function of the difference in the variance components (figure 4). Eventually, for some value of θ_U greater than 2.0, there is no difference between the adjusted versus the unadjusted power. The point at which there is no further difference between these powers (θ_U^*) is a function of the overall prevalence and the stratum sample fractions (figure 5).

This investigation has important implications for the evaluation of sample size a priori for clinical trials and epidemiologic investigations. In a randomized clinical trial, randomization on the whole protects against a severe covariate imbalance. Thus, any difference between the unadjusted and the adjusted odds ratio should be negligible. Although other covariates may be strongly associated with the outcome, this alone will have a trivial effect on power if there is no covariate imbalance to accompany it. Further, the power is largely unaffected by the presence of heterogeneity of treatment effect among strata. Thus, for a randomized clinical trial, the sample size and power can be accurately determined on the basis of an unstratified assessment.

For a nonrandomized study, however, the situation is entirely different. Again, heterogeneity among strata will not materially affect power. However, in a nonrandomized study, a covariate imbalance or confounding is not only possible but also a major concern. In this case, a stratified assessment of power that accounts for the potential for both a covariate imbalance and covariate association is preferred. However, if there are only two strata and the sample fraction in one stratum is very small (less than 0.1) and the prevalence of response is close to 0.50, then there will be a negligible difference between the adjusted and the unadjusted power if one wishes to detect an unadjusted odds ratio in the order of 2.2 or greater.

Appendix A. Constructing the pooled table given N, Qe, $\pi_{\bullet+\bullet}$, and θ_u

The row and column totals of the pooled table are obtained as follows:

$$n_{\bullet\bullet\bullet} = N \quad n_{e\bullet\bullet} = Qe \cdot N; \qquad n_{c\bullet\bullet} = N - n_{e\bullet\bullet}$$
$$n_{\bullet+\bullet} = \pi_{\bullet+\bullet} \cdot N; \quad n_{\bullet-\bullet} = N - n_{\bullet+\bullet}$$

Given θ_U, the cell (1,1) entry $n_{e+\bullet}$ is obtained by solving the equation

$$\theta_U = \frac{n_{e+\bullet}([1 - Qe - \pi_{\bullet+\bullet}]N + n_{e+\bullet})}{(QeN - n_{e+\bullet})(\pi_{\bullet+\bullet}\, N - n_{e+\bullet})} \tag{A.1}$$

for $n_{e+\bullet}$. When $\theta_U = 1$, then $n_{e+\bullet} = N \cdot Qe \cdot \pi_{\bullet+\bullet}$. When $\theta_U \neq 1$, equation (A.1) is quadratic in $n_{e+\bullet}$. In this case, choose the positive root (rounded to the nearest whole number) that best approximates equation (A.1) as the cell (1,1) entry $n_{e+\bullet}$. After $n_{e+\bullet}$ and the row and column totals are obtained, the rest of the entries automatically follow.

Appendix B. Constructing the stratum-by-group table given r_1 and θ_I

$$n_{\bullet\bullet1} = r_1 \cdot N; \quad n_{\bullet\bullet2} = N - n_{\bullet\bullet1}$$

The column totals $n_{e\bullet\bullet}$ and $n_{c\bullet\bullet}$ have already been obtained in appendix A. Given θ_I, the cell (1,1) entry $n_{e\bullet1}$ is obtained analogous to how $n_{e+\bullet}$ is obtained. In equation (A.1), substitute θ_I for θ_U, $n_{e\bullet1}$ for $n_{e+\bullet}$, and r_1 for $\pi_{\bullet+\bullet}$; then solve for $n_{e\bullet1}$. The rest of the table entries follow after $n_{e\bullet1}$ is obtained.

Appendix C. Constructing the stratum-by-response table given θ_A

The row and column totals have already been obtained in appendices A and B. The cell (1,1) entry $n_{\bullet+1}$ is obtained as follows: In equation (A.1), substitute θ_A for θ_U, $\pi_{\bullet+\bullet}$ for Qe, r_1 for $\pi_{\bullet+\bullet}$, and $n_{\bullet+1}$ for $n_{e+\bullet}$; then slove for $n_{\bullet+1}$. The rest of the table entries follow after $n_{\bullet+1}$ is obtained.

Appendix D. Constructing the stratum 1 and stratum 2 tables given ψ

Stratum 1 and stratum 2 tables can be expressed in terms of the entries of the pooled, stratum-by-group, and stratum-by-response tables as follows:

S1	e	c	Total
+	n_{e+1}	$n_{\bullet+1} - n_{e+1}$	$n_{\bullet+1}$
−	$n_{e\bullet1} - n_{e+1}$	$n_{c\bullet1} - n_{\bullet+1} + n_{e+1}$	$n_{\bullet-1}$
Total	$n_{e\bullet1}$	$n_{c\bullet1}$	$n_{\bullet\bullet1}$

S2	e	c	Total
+	$n_{e+\bullet} - n_{e+1}$	$n_{\bullet+2} - n_{e+\bullet} + n_{e+1}$	$n_{\bullet+2}$
−	$n_{e\bullet2} - n_{e+\bullet} + n_{e+1}$	$n_{c\bullet2} - n_{\bullet+2} + n_{e+\bullet} - n_{e+1}$	$n_{\bullet-2}$
Total	$n_{e\bullet2}$	$n_{c\bullet2}$	$n_{\bullet\bullet2}$

After constructing the pooled, stratum-by-group, and stratum-by-response tables, only the cell (1,1) entry of the stratum 1 table, n_{e+1}, is needed in order to completely determine the two tables. Given ψ, n_{e+1} is obtained by solving the equation

$$\psi = \frac{\begin{array}{c} n_{e+1}\,(n_{c\bullet 1} - n_{\bullet +1} + n_{e+1})(n_{e\bullet 2} - n_{e+\bullet} + n_{e+1}) \\ (n_{\bullet +2} - n_{e+\bullet} + n_{e+1}) \end{array}}{\begin{array}{c} (n_{e\bullet 1} - n_{e+1})\,(n_{\bullet +1} - n_{e+1})\,(n_{e+\bullet} - n_{e+1}) \\ (n_{c\bullet 2} - n_{\bullet +2} + n_{e+\bullet} - n_{e+1}) \end{array}} \tag{D.1}$$

for n_{e+1}. Equation (D.1) is a third-degree polynomial in n_{e+1} when $\psi = 1$ and a fourth-degree polynomial in n_{e+1} when $\psi \neq 1$. Pick the positive real root of equation (D.1) (rounded to the nearest whole number) as the cell (1,1) entry of the stratum 1 table. If more than one positive real root exist, pick the root that, when rounded to te nearest whole number, best approximates equation (D.1). Once n_{e+1} is specified, the rest of the table entries follow.

Appendix E. Specified odds ratios versus attained odds ratios

Given a specified unadjusted odds ratio, say $\theta_{U(\text{specified})}$, the solution $n_{e+\bullet}$ obtained from equation (A.1) may or may not be integer valued. In constructing the pooled 2 × 2 table, the solution $n_{e+\bullet}$ obtained is rounded to the nearest integer. The resulting odds ratio from the constructed 2 × 2 table, say $\theta_{U(\text{attained})}$, may or may not be exactly the same as $\theta_{U(\text{specified})}$ due to the rounding of the solution $n_{e+\bullet}$ of equation (A.1).

The same is true for the constructed stratum-by-group, stratum-by-response, stratum 1, and stratum 2 tables. Hence there may be distinctions between $\theta_{I(\text{specified})}$ and $\theta_{I(\text{attained})}$, $\theta_{A(\text{specified})}$ and $\theta_{A(\text{attained})}$, and $\psi_{(\text{specified})}$ and $\psi_{(\text{attained})}$.

References

1. Halperin M, Rogot E, Gurian J, Ederer F (1968). Sample sizes for medical trials with special reference to long-term therapy. *J Chron Dis* 21:13–24.
2. Lachin JM (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clin Trials* 2:93–113.
3. Donner A (1984). Approaches to sample size estimation in the design of clinical trials — a review. *Stat Med* 3:199–214.
4. Mantel N, Haenszel W (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J Nat Cancer Inst* 22:719–748.
5. Cochran WG (1954). Some methods of strengthening the common χ^2 test. *Biometrics* 10:417–451.
6. Radakhrishna S (1965). Combination of results from several 2 × 2 contingency tables. *Biometrics* 21:86–98.
7. Birch MW (1964). The detection of partial association. I. The 2 × 2 case. *J R Stat Soc B* 26:313–324.
8. Woolson RF, Bean JA, Rojas PB (1986). Sample size for case-control studies using Cochran's statistic. *Biometrics* 42:927–932.

9. Wittes J, Wallenstein S (1987). The power of the Mantel–Haenszel test. *J Am Stat Assoc* 82:1104–1109.
10. Hardison CD, Quade D, Langston RD (1986). Nine functions for probability distributions. In *SUGI supplemental Library User's Guide, Version 5 Edition*, RP Hastings (ed.). Cary, NC: SAS Institute, 385–393.
11. Rochon J (1989). Application of the GSK method to determination of minimum sample sizes. *Biometrics* 45:193–205.
12. Whittemore A (1981). Sample size for logistic regression with small response probability. *J Am Stat Associ* 76:27–32.
13. Hsieh EY (1989). Sample size tables for logistic regression. *Stat Med* 8:795–802.
14. Wilson SR, Gordon I (1986). Calculating sample size in the presence of confounding variables. *Appl Stat* 35:207–213.
15. Self SG, Mauritsen RH (1988). Power/sample size calculations for generalized linear models. *Biometrics* 44:79–86.
16. Beach ML, Meier P (1989). Choosing covariates in the analysis of clinical trials. *Controlled Clin Trials* 10:161S–175S.
17. Canner PL (1991). Covariate adjustment of treatment effects in clinical trials. *Controlled Clin Trials* 12:359–366.
18. Canner PL (1981). Choice of covariates in the adjustment of treatment effects. Presented at the Society for Clinical Trials Annual Scientific Sessions, San Francisco.
19. Greenland S (1985). Power, sample size, and smallest detectable effect determination for multivariate studies. *Stat Med* 4:117–127.

11. Quality-of-life assessment in clinical trials

Richard D. Gelber and Shari Gelber

Introduction

Within the context of clinical trials, quality of life (QOL) is a multidimensional concept that encompasses health-related constructs, but excludes other dimensions such as economics, housing, or education. Most QOL research has been based on the World Health Organization (WHO) definition of health [1]: 'Health is not only the absence of infirmity and disease but also a state of physical, mental and social well-being.' Thus QOL encompasses all health-related outcomes beyond those of survival and physiological responses. Diseases and their treatments affect not only patients' physical functioning and level of pain, but also their cognitive, emotional, and social functioning. QOL measures have also included assessments of sexual functioning, family and marital relationships, role performance, vitality, sleep, health perceptions, general life satisfaction, and symptoms such as nausea and fatigue. QOL assessment has been employed in developing individual patient treatment plans, performing cost–benefit analyses, making health policy decisions, and conducting clinical trial evaluations. Health status, functional status, and health-related quality of life have become synonyms for QOL in the clinical trials literature.

This chapter will focus on methodological issues pertaining to the use of QOL assessments in the evaluation of clinical trial results. We will first discuss the purpose of QOL in clinical trials and then summarize the historical background. Next we provide a detailed overview of the available instruments and the standards for their selection. Then we focus on the special features of the design and conduct of clinical trials that incorporate QOL assessments, and we also outline the criteria for selecting a statistical methodology. Finally, we review some of these statistical approaches.

The purpose of QOL in clinical trials

The increasing importance of a QOL assessment has become well recognized [2,3]. Medical therapies sometimes compromise patients' physical, psycho-

Peter F. Thall (ed.), RECENT ADVANCES IN CLINICAL TRIAL DESIGN AND ANALYSIS.
Copyright © 1995 Kluwer Academic Publishers, Boston. All rights reserved. ISBN 978-1-4613-5830-5.

logical, and social functioning. If a particular treatment offers a minimal gain in survival, it is especially important to assess the net benefit after the toxicity is considered. These QOL considerations may differ among the prognostic subgroups or may change over time. The joint working party of the Food and Drug Administration and the National Cancer Institute has developed recommendations for treatment endpoints in QOL [4].

There are four types of clinical trials in which the use of QOL assessments as clinical outcomes is crucial [5]. First are trials in which the treatment is not expected to alter the course of the disease, but instead to provide symptom relief. Then there are the trials of potentially toxic drugs expected to decrease disease morbidity and mortality. Third are prevention trials that assess interventions on asymptomatic participants. Finally, there are comparisons of new drugs that are hoped to be less costly or to have fewer side effects than the standard therapy.

Coates et al. [6] used QOL assessments as predictors of outcome rather than as endpoints in a clinical trial. Measurements at baseline of these predictive factors may allow clinicians to better select appropriate treatments for their patients.

Historical background

Karnofsky is generally credited with introducing the first measure of physical functioning in 1948 [7]. The Karnofsky Performance Status evaluates patients' physical daily functioning on an 11-point scale from 0% (death) to 100% (completely normal), representing 'approximate percentage of normal physical performance.' This measure has been used extensively in cancer clinical trial evaluations, despite the fact that its psychometric properties were not formally analyzed until 1980 [8]. In 1960 the Eastern Cooperative Oncology Group (ECOG) reduced the Karnofsky index into a six-point performance scale, often referred to as the Zubrod scale [9].

In 1949 Steinbroker et al. introduced an index for measuring functional status in rheumatoid arthritis [10], and in 1964 The Criteria Committee of the New York Heart Association recommended a classification system for patients with cardiac disease, utilizing four functional classes and five therapeutic categories [11]. In 1975 Patterson [12] suggested a three-level survival quality index based on the duration of the clinical response, symptomatic or functional QOL impairment, usefulness of the response, and cost of the treatment. Thus the earliest evaluations of QOL identified and quantified the physical effects of disease and its treatment. Subsequent measures have incorporated the patients' perspectives of their illnesses and therapeutic regimens.

The medical literature contains several early clinical trials that included multidimensional QOL measurements. In 1971 Izsak and Medalie [13] developed a multidimensional scale that measured physical, social, and psychological variables in cancer patients. The scale was tailored to specific cancers and designed to assist clinicians in determining rehabilitation needs and evaluating patient progress. In 1975 a trial for patients with acute myelogenous leukemia used a six-level assessment of QOL ranging from 'hospital stay throughout illness' to 'no symptoms, normal life' [14]. The assessments were based on patient reports of their symptoms and functioning. Given the usefulness of these early studies, one wonders why they did not stimulate other investigators to include QOL measures routinely in clinical trials. This may be due in part to the difficulty of dealing simultaneously with both QOL and response measures. Researchers tend to be more comfortable with 'objective' endpoints than with 'subjective' measures.

The modern era of QOL in cancer clinical trials research is generally cited to have begun in 1976 with Priestman and Baum's study of breast cancer treatment [15]. They used a linear analogue self-assessment scale to measure QOL, with 10 questions assessing general feeling of well-being, mood, level of activity, pain, nausea, appetite, ability to perform housework, social activities, general level of anxiety, and overall treatment assessment. Their results indicated that this instrument could be used to assess the subjective benefit of treatment in individual women, to detect changes over time, and also to compare different treatments within a clinical trial.

Instruments

Selection

During the past two decades, numerous instruments have been developed and successfully used in the evaluation of cancer treatments in clinical trials. There are several issues that need to be considered when selecting an appropriate instrument for assessing QOL within a clinical trial: (1) the purpose of the clinical trial, (2) the patient population, (3) the treatments and their potential toxicities, and (4) the resources of the investigators and the participating clinicians. These issues will determine the type of instrument selected and the method of administration. For example, if the disease being studied or its treatment toxicities severely compromise functioning, an instrument should be selected that requires minimal effort on the part of the patient.

Global and specific functioning

The purpose for which the assessment will be used within the clinical trial will determine which domains will be measured and whether the assessment will be of global or specific functioning, or of both. For the purpose of clinical trial research, the most frequently assessed domains are physical, cognitive, and social functioning, patient satisfaction, and emotional well-being. A global measure provides a single composite score that encompasses multiple domains. Specific measures provide data on one or more symptoms or on a single domain.

Generic instruments

A second classification of instruments is generic versus disease specific. Generic instruments assess general domains of health rather than particular symptoms. They permit comparisons among different diseases and between different patient groups. Since generic measures are developed to assess general populations, they are less sensitive to the small but clinically significant differences that may be relevant when comparing treatment arms for a specific disease setting.

Generic instruments can be either health profiles or utility measures. A health profile evaluates the major domains of general health. Examples of health profiles are the Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36) [16], Spitzer's Quality of Life-Index (QL-Index) [17], and the Quality of Well-Being (QWB) Scale [18].

Utility measures were developed in the fields of economics and decision science. They provide a numeric measure for patients' assessments of QOL outcomes on a scale of 0.0 (death) to 1.0 (perfect health). Since utilities are single measures of a health state, they can be used to evaluate overall changes in patients, but do not indicate in which domains these changes have occurred. A frequently used method for estimating utilities is the 'time trade-off' (TTO) [19], in which patients choose between a length of life in their current state of health and a lesser lifetime in perfect health. Utilities are used to calculate quality-adjusted life-years (QALYs), which can be used to determine quality-adjusted life expectancies [20].

Disease-specific instruments

Disease-specific measures are designed to evaluate outcomes in a particular disease and thus are narrower in focus. These instruments are less likely to be generalizable to other patient populations. They generally measure areas that are of greatest concern to the patients and clinicians and are more sensitive to the relevant differences in the comparison of therapies. The Breast Cancer Chemotherapy Questionnaire (BCQ) [21] is an example of a disease- and treatment-specific instrument.

To ensure that the assessment is comprehensive enough to measure all possible side effects, it may be necessary to create a new disease-specific questionnaire. However, it is very time-consuming to validate a new instrument and establish its reliability and responsiveness to the clinically relevant differences. Frequently a test battery, combining generic and disease-specific measures, is used to provide a more comprehensive evaluation of patient responses. For example, the Functional Assessment of Cancer Therapy scale (FACT) [22] consists of a core set of generic questions and disease-specific modules.

Instrument administration

Another decision is whether the assessment is obtained directly from the patients or provided by family members or health care professionals. Changes in clinical status assessed by clinicians may not provide good indications about how the patients are functioning or feeling. Research has shown poor correlation between patient and clinician responses [1]. Part of this discrepancy is due to the patients' ability to accommodate to or cope with the limitations imposed by their illness.

There are three main methods for collecting QOL data: self-administered questionnaires, telephone interviews, and face-to-face evaluations. Self-administered instruments are less expensive, but must be constructed with careful attention to reading levels, clarity, length, and patient characteristics. A brief questionnaire will increase respondent compliance. Telephone and face-to-face interviews are more costly due to the administration time and the need to train the interviewers. However, they generally produce better data quality than self-administered instruments and are more accessible to patients who have limited physical functioning. Self-administered questionnaires are most often employed in the clinical trial setting because they are less expensive and more efficient than the other forms of data collection. Trials are underway in which patients use computers to directly enter their responses to QOL questions [23]. This may improve the data quality for those patients who can utilize this method.

Item selection

Questions can be asked either in closed or open-ended form. Open-ended questions permit the patients to provide more personal responses, but these are difficult to code and quantify. Thus the closed form is most commonly used. Questions can be in the form of yes/no or true/false or on a scale. These scales are generally divided into two categories: the Likert, a categorical scale, and the continuous visual or Linear Analogue Self Assessment (LASA) scale. The Likert scale provides the patient with a limited choice of clearly defined response categories. The labeling of these categories can sometimes lead to confusion when the ordering appears to be out of sequence. The number of categories will influence the patients' answers; too many can be confusing, and too few may be too restrictive. There is some disagreement as to whether to include a middle or neutral response category [24]. The most frequently employed scales use either four or five categories [25].

In contrast to the Likert scale, the analogue scale is an unmarked line, generally 10 cm long, which is anchored on either end with antonym adjectives or descriptions. It is most often used for subjective measures such as mood. The respondents place a mark anywhere along the line that best reflects their answer. Although the true linear analogue scale provides for fine discrimination, its administration has several practical limitations. It requires an extra step for the data manager to measure and code each response for analysis.

There are many advantages to using multiple items to measure a particular construct. Multiple items will make the instrument more sensitive to small treatment differences. The stability of responses also increases when more than one question is used to create a summary score. The validity of the scale is improved when the group of items are carefully selected to be representative of the attribute being measured. Missing responses can also be replaced by a summary score of the completed answers in the same area.

These advantages must be weighed against the additional resources required to administer and analyze lengthy, multiple-item scales. These longer instruments also increase the time and energy needed by the patients to complete the questionnaire.

Cross-cultural studies

Instruments to be used in cross-cultural studies have additional criteria that require special attention. Different cultures often have different attitudes toward concepts such as illness and pain [26]. Different cultural beliefs about privacy will impact on the responses patients provide. Willingness to admit feeling pain will also vary.

The first step in developing instruments to be used cross-culturally is to translate the questionnaires. This should be done by native speakers of the second language and 'back-translated.' The translated instrument must then be piloted to assure that the translation maintains the intent of the original instrument. Cultural differences may still strongly influence the meaning of the responses, despite an accurate translation. The anchoring of Likert scales is especially difficult to reproduce in a new language, or even with the same language but in a different cultural context. Measures should always be normed when applying them in different cultures. In one norming strategy, the results obtained from administering the new instrument to two different patient populations are compared to the differences that would have been expected based on known influences or characteristics of the two subgroups.

Psychometric properties

The instrument must have been proven reliable, valid, and responsive in comparable populations. Guyatt, Feeny, and Patrick [27], Stewart [28], and

Deyo, Diehr, and Patrick [29] provide detailed discussions of these concepts. Reliability refers to the stability of the response under the same stimulus from one occasion to another. Reliability should be measured using one or more of the following: test–retest, internal consistency (Cronbach's alpha) [30], and interrater reliability. Validity refers to the ability of the instrument to measure the constructs that it purports to measure. There are no gold standards against which health outcome assessments can be compared. It is therefore necessary to use content, criterion, and construct validity. Content validity is the association between the scale and all the appropriate theoretical domains. Criterion validity is the extent to which the instrument's responses correspond to those of a criterion standard. And construct validity is the relationship between the results of the measure and what would be expected from particular patient groups. In addition, in clinical trials the measure must be shown to be responsive to the clinically important differences between the treatments being compared. Responsiveness is influenced by the number of items pertaining to each area being measured and the number of response category levels per item. For example, one yes/no question concerning physical functioning will be much less responsive then six items rated on a five-point Likert scale.

Examples of instruments used in cancer clinical trials

Some of the most frequently used instruments in cancer clinical trials are Karnofsky Performance Status [7], ECOG (Zubrod) [9], Breast Cancer Chemotherapy Questionnaire (BCQ) [21], Cancer Rehabilitation Evaluation System (CARES) [31], European Organization for Research and Treatment of Cancer (EORTC) scale [32], Functional Assessment of Cancer Therapy (FACT) [22], Functional Living Index–Cancer (FLIC) [33], International Breast Cancer Study Group Quality of Life questionnaire (IBCSG–QL) [34], Linear Analogue Self Assessment (LASA)–Priestman and Baum [15], QLI–Coates [6], Quality of Life Index (QL-Index) [17], and the MOS 36-Item Short-Form Health Survey (SF-36) [16]. An annotated list of these instruments is presented in table 1.

Design and conduct of clinical trials

Clinical trials that evaluate QOL endpoints must adhere to the rigorous design criteria required for any effective study. The need for randomization and double blinding is especially crucial when measuring subjective outcomes. The sample size must be large enough to permit detection of clinically significant differences. However, QOL issues can often be studied with sample sizes smaller than those required for the primary efficacy analysis. The power to detect differences in QOL is often increased by using methods for longitudinal data analysis — for example, repeated measures [35].

Table 1. Instruments frequently used in cancer clinical trials

Instrument	Number of items	Scale	Domains assessed
Karnofsky Performance Status	1	Likert	Physical
ECOG (Zubrod) Scale	1	Likert	Physical
BCQ	30	Likert	Attractiveness, fatigue, physical symptoms, inconvenience, emotional, hope, social support
CARES	93–132	Likert	Physical, psychosocial, medical interaction, marital, sexual, symptom-and treatment-specific items
EORTC: QLQ–C30	42	Likert/ binary	Five functional scales (physical, role, cognitive, emotional, and social), three symptom scales (fatigue, pain, nausea) with additional disease-specific items and a global QOL scale
FACT	36–40	Likert	Physical, social/family, relationship with doctor, emotional, functional, well-being, and additional disease-specific items
FLIC	22	Analogue	Psychological, social, disease symptoms, global well-being, treatment and disease issues, and physical functioning
IBCSG–QL	10	Analogue	Physical well-being, mood, fatigue, appetite, coping, social support, symptoms, overall health
LASA–Priestman and Baum	25	Analogue	Physical, psychological, social
QLI–Coates	4–5	Analogue	Physical, psychological
QL-Index	5	Likert	Physical activity, daily living, health perceptions, psychological, social support, outlook on life
SF-36	36	Likert	Physical functioning, role physical, pain, general health perceptions, vitality, social functioning, role emotional, mental health, well-being

Scheduling of assessments

The schedule and frequency of QOL assessments should be defined to optimally detect treatment differences. The timing should reflect anticipated disease progression and treatment toxicities. All treatment arms must be assessed at the same time points. The specifics of administration of the measure at each visit might also influence the response. For instance, is the questionnaire completed before, during, or at the end of the clinic visit? Is the instrument presented to the patient by the doctor, by the nurse, or by other clinic staff? If they are administered too often, patients may refuse to complete the questionnaires, and the clinic staff may find the effort too burdensome.

In general, there should be a minimum of three administrations: one at baseline, once when the treatment toxicities are the most prominent, and finally one at the end of treatment when the therapeutic benefit is most likely to be present. A baseline measurement that precedes therapy allows for assessment of treatment-related changes within an individual patient. Measurements during the therapeutic regimen permit patients to report pertinent side effects. It must be clear to the patients whether they should describe their experiences of the effects of the therapy, the rest periods, or both. The specific time period to which each question refers (e.g., today, within the last two weeks, or since the last clinic visit) must also be specified. It is important to have a sufficiently long period of follow-up to allow for the treatment effect and potential late sequelae to be recorded.

Data quality and completeness

Obtaining complete data from patients is critical inorder to achieve an accurate comparison of the treatment arms. To avoid problems with data collection, it is important to designate one or more individuals at each site to be responsible for completion of the froms. The importance of the trial and of the infromation that they are providing should be stressed to all patients. If possible, it is also useful to obtain measurements on those patients who discontinue the protocol study. For the QOL assessment to be successful, it must be incorporated as a routine function within the clinic visit, having the same status as laboratory studies for clinical monitoring of toxicities and disease progression.

Criteria for selecting statistical methodology

There are a number of statistical methodologies employed in the analysis of QOL data. Each method requires specific assumptions and yields different summary measures. The analyses should be understood by clinicians so they can utilize the trial results in their practices. The reporting and interpreting of the results must be clear and concise and must highlight clinically relevant differences. Most QOL questionnaires generate multiple measurements on one or more dimensions. Some instruments are designed to provide summary scores for particular domains, but often multiple domains are assessed. It would be difficult to base a clinical decision on conflicting conclusions from numerous QOL factors in a single study. It is particularly useful to incorporate survival data into the QOL analysis. The evaluation of different treatment arms requires the assessment of quantity as well as quality gains.

The analysis should be able to handle missing data. Infromation is frequently lost due to adverse outcomes such as death or disease progression that prevent the collection of QOL data. Therefore, the missing data are not missing at random but rather are informatively censored, and an adjustment must be made in the analysis. Finally, the treatment comparison should include a sensitivity analysis so that patient preferences can be incorporated. Individual patients will have a variety of approaches to the trade-offs that are often relevant when making therapeutic choices.

Statistical analysis

Statistical methods used for the analysis of QOL data have been reviewed by Fayers and Jones [25], Olschewski and Schumacher [36], Schumacher et al. [37], and Cox et al. [38]. Many of the usual methods of data analysis are applicable. In this section, we discuss some of the statistical methods developed specifically for QOL clinical trial treatment comparisons (see table 2).

Descriptive statistics

A descriptive summary of the data is most often used to evaluate QOL assessments. Schumacher, Olschewski, and Schulgen [37] reviewed articles published between 1985 and 1989 and found that about one third used only descriptive statistics to analyze the data. The QOL results were reported as frequencies, mean differences, and correlations between different item responses, or between a response and a clinical measure.

When the data collected are continuous (for example, from an analogue scale), then the results may be presented as a distribution with means and standard errors reported. Patient–group mean changes over time may also

Table 2. Some methods for QOL statistical analysis

Descriptive
Means, standard errors, frequencies, correlations, graphs
Analytic—univariate
t-tests, ANOVA, nonparametric methods
Life table analyses of QOL endpoints
Analytic—multivariate
Repeated measures ANOVA
MANOVA
Mixed effects models
Growth curve analysis
Global statistics
Analytic multivariate, incorporating survival data
Markov and semi-Markov models
Q-TWiST

be plotted. Very often the QOL scores on a continuous scale cluster toward the extremes or at discrete points along the scale.

Categorical scales (e.g., Likert scales) are also used for QOL assessments. The distributions of these scores should be examined and nonparametric methods used if the assumption of normality is violated. Descriptive results from either discrete or continuous scales can be compared over time and reported graphically, as presented by Nou and Aberg [39].

Analytic univariate methods

The second general category is an analytic approach using either a univariate or multivariate analysis. In a univariate analysis, no attempt is made to jointly analyze the multiple endpoints. Each endopint is analyzed and reported separately. The review of articles published between 1985 and 1989 revealed that 50% of the articles used univariate analyses [37].

QOL assessments can be used to define an occurrence of an endpoint. For example, deterioration below a specified level of minimal functioning can be used to indicate treatment failure. Life table methods can then be applied to estimate the risk of failure and to compare treatments with respect to the time to reach this QOL-oriented endpoint [40]. Rosenman and Choi [41] are credited with the first application of a Kaplan–Meier analysis for an endpoint based on QOL (alive but with Karnofsky Index less than 60).

Univariate treatment comparisons are made on a single factor to determine if one group has better QOL than the others. If the underlying distribution of the QOL parameter, or its transform, is assumed to be normal, a *t*-test or ANOVA may be used for the comparison. Otherwise, a nonparametric approach should be used. The multiple comparison problem is the major disadvantage of univariate analyses. Multiple endpoints analyzed separately can lead to false positives. The Bonferroni method can be used to protect the type 1 error, but this represents a very conservative approach when there are a large number of endpoints.

Analytic multivariate methods

Treatment groups within a clinical trial may be compared using repeated measures over time (longitudinal analysis), multiple endpoints of an individual patient at a single point in time, or multiple endpoints gathered at several time points. Multivariate techniques include repeated measures ANOVA, MANOVA (multivariate analysis of variance), mixed effects modeling, and growth curve analysis. While these provide general comparisons of treatments, other methods have been proposed specifically to use multivariate QOL data.

Global statistics. One approach to the analysis of multiple endpoints is the use of global statistics [42,43]. A treatment comparison is performed for each of the multiple items, and the results are combined to form a single global statistic. O'Brien [42] developed both parametric (using *t*-tests) and nonparametric (using Rank Sum tests) approaches employing a weighted combination of the individual test statistics using the covariance matrix for the derivation of the weights. Pocock et al. [43] extended the approach to any set of asymptotically normal individual test statistics. Tandon [44] applied global statistics to QOL data. The usual multivariate analysis tests the hypothesis that treatments are different, without requiring that the direction of the effect be the same for each dimension. Global statistics, however, test the more relevant hypothesis: is one treatment consistently better than the other based on each of multiple dimensions assessed in the QOL instrument? Global statistics are especially useful when the sample size is small and the QOL instrument has a large number of closely related measures.

Repeated measures. An example of repeated-measures modeling was proposed by Zwinderman [45] to assess QOL when comparing two or more treatment groups in a clinical trial. The QOL data were multidimensional dichotomous variables measured at different time points. He used a latent-trait stratified logistic regression model with parameters for the time process, the treatment effect, and the interaction between the two. The latent trait is a factor, not directly observable, that is estimated through its theoretical link with a set of measurable behaviors.

Under this model, if the interaction is not statistically significant, then the time process and treatment effects can be estimated independently of the baseline QOL measurements. Thus it is possible to ignore differential mortality, censoring mechanisms, and other missing data problems. The key assumption is that the lack of an interaction implies that patients' treatment response is not dependent upon their baseline QOL measures. Zwinderman comments that the model is biased when a trade-off exists between QOL and mortality and suggests performing a survival analysis concurrently.

Markov and semi-Markov models. Markov and semi-Markov models have been used to compare treatments based on estimates of the time spent in different health states and the probabilities of transitions between these states. The relevant health states must be identified, and then each is weighted to reflect the relative value of a health state compared to perfect health. The treatments are then compared in terms of the total quality-adjusted time, i.e., the weighted sum of the health state durations.

In general, to calculate the transition probabilities, an underlying model must be assumed. The most commonly used model is the Markov chain, which assumes that the transitions from one QOL state to another are independent and continuous and only depend upon the previous state. This requires that the assessments are made at time points independent of the patients' treatment schedule or health state.

Several investigators have used these stochastic models to evaluate QOL data. Gore [46] used health state transitions to create a 'grid of quality-adjustments' to evaluate trade-offs between different radiation schedules for cancer. This grid is a chart of logrank tests comparing the quality-adjusted survival times for varying utility weights assigned to different health states.

Loewy et al. [47] used survival data to distinguish two attributes of utility, namely, prolongation of life and QOL, for evaluating beta-blockers for heart disease. Discrete-time transient semi-Markov processes were used to model the health state transition probabilities corresponding to the prolongation of life, while a simple recurrent Markov process was used to derive the QOL state transition probabilities. In a semi-Markov process, the state changes form an embedded Markov chain, and the times spent in different health states are mutually independent and depend only on the adjoining states [48]. Researchers have used Markov models to study systemic lupus ertematosus [49], recurrences of headaches [50], prostate cancer [51], and hepatocellular carcinoma [52].

An alternative approach proposed by Pepe [53] uses a qualifier Q for the survival function that estimates the prevalence of a transient condition among surviving patients over time. This nonparametric estimator can be used in conjunction with the survival curves to describe a major aspect of QOL. It was applied to estimate the prevalence of chronic graft-versus-host disease among patients receiving bone marrow transplantation and was contrasted with Markov and semi-Markov estimates using the same data set.

Q-TWiST methodology. The objective of the Q-TWiST method is to incorporate QOL aspects into treatment comparisons in clinical trials so that the evaluation of therapies is based on both quantity and quality of life. The starting point is the overall Kaplan–Meier survival curve, which is often used as the primary endpoint in clinical trials. Q-TWiST can be used not only in treatment evaluations, but also to facilitate individual therapy choices incorporating patient preferences regarding treatment toxicity and efficacy. Q-TWiST stands for Quality-adjusted *T*ime *Wi*thout *S*ymptoms of disease and *T*oxicity of treatment [54]. It is based on the concept of quality-adjusted life-years (QALYs) [20] and represents a utility-based approach to QOL assessment in clinical trials [55].

The development of the method was motivated by a medical controversy in which the study of adjuvant therapy in breast cancer patients demonstrated an improved disease-free survival but not overall survival [56,57]. Q-TWiST provided an opportunity to balance the improved QOL associated with delayed recurrence against the diminished QOL associated with toxic side effects of the adjuvant therapy. The technical aspects of the method were described by Goldhirsch et al. [54] and Glasziou et al. [58]. The method has been applied to address treatment comparisons for patients with node-positive [59] and node-negative breast cancer [60] as well as patients with HIV infection [61,62].

METHODOLOGY. The application of the Q-TWiST method involves the following three steps.

1. The first step is to *define* QOL-oriented health states that highlight the differences between the regimens being compared for the specific disease under study. Included among these states is TWiST, a period of relatively uncompromised QOL, representing the best QOL available for the study patients. Patients progress through the health states chronologically, possibly skipping one or more states, but never backtracking. These states can be defined retrospectively at the time of data analysis or can be specified prospectively in the protocol document in anticipation of performing a Q-TWiST analysis.

2. In the second step, Kaplan–Meier curves for health state transition times are used to *partition* the area under the overall survival curves separately for each treatment. The average time a patient spends in each health state through the follow-up period is calculated using restricted means [62]. These are restricted because they represent the mean health-state duration up to a finite time determined by the follow-up interval for the study. For example, if the median follow-up of patients in a breast cancer study was seven years, than it would be reasonable to calculate estimates of time spent in each health state within seven years from randomization (i.e., average times restricted to seven years.)

3. The third step is to *compare* the treatment regimens using the weighted sum of the mean durations of each health state as calculated in step 2. This quality-adjusted survival comparison offers the opportunity to include utility weights to reflect the assumed relative value to the patient of the different health states. On a scale of 1.0 to 0.0, these utility coefficients represent the worth of the different states, where TWiST is assigned the value of 1.0 and death is assigned the value of 0.0. Thus a weight of 0.5 reflects a health state that is valued half as much as time spent in TWiST. Treatment comparisons can be made using sensitivity analyses, also called threshold utility analyses.

AN EXAMPLE. Lenderking et al. [62] recently published a Q-TWiST analysis to evaluate the QOL associated with zidovudine treatment in asymptomatic HIV-infected patients. The original study performed by the AIDS Clinical Trials Group (ACTG 019) was a double-blind, randomized, placebo-controlled clinical trial designed to study the efficacy and safety of two different doses of zidovudine (1500 mg and 500 mg daily) in asymptomatic HIV-infected patients [64]. The Q-TWiST analysis evaluated the trade-off between the potential delay in disease progression and the toxicities associated with the use of the two doses of zidovudine in patients who were experiencing little or no disease-related symptoms at the time of study entry.

In step 1 of the Q-TWiST analysis, the following three health states were identified: (1) TWiST, the number of months preceding the development of a grade 3 or worse symptomatic adverse event or HIV disease progression, whichever occurred first; (2) AE, the period after the first occurrence of a severe adverse event; and (3) Prog, the period after the progression of the HIV disease. The Q-TWiST survival model used the utility coefficients u_{AE} and u_{Prog} to reflect the value of these states of health (time after adverse event and time after disease progression). TWiST was assigned a weight of 1.0 and death a weight of 0.0. Thus the QOL-adjusted survival relative to TWiST (Q-TWiST) was calculated as:

$$\text{Q-TWiST} = \text{TWiST} + (u_{\mathrm{AE}} * \text{AE}) + (u_{\mathrm{Prog}} * \text{Prog})$$

Figure 1 shows an example of the different time periods with arbitrary utility coefficients of 0.75 for AE and 0.5 for Prog.

Step 2 used the collected data to calculate separate Kaplan–Meier curves for the adverse event-free survival (AEFS), progression-free survival (PFS), and overall survival (OS). These curves partition the OS into periods of time in TWiST, time after an adverse event, and time after disease progression. The partitioned survival analyses for the standard-dose zidovudine group (500 mg daily) and for the placebo group are shown in figure 2. The results were restricted to the first 18 months of follow-up. Patients treated with zidovudine therapy spent less time in Prog, but more time in AE than patients in the placebo group. The two groups had approximately equal amounts of TWiST.

In step 3, the two treatment groups (zidovudine 500 mg daily versus placebo) were compared using a threshold utility analysis of the amount of

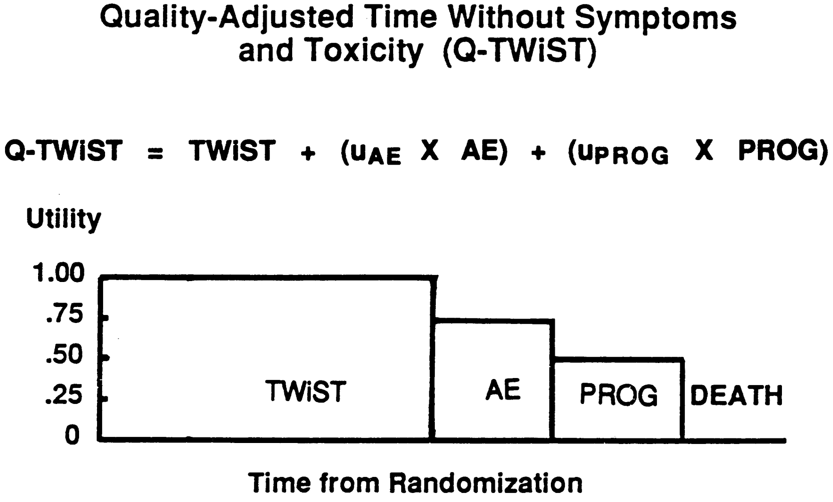

Quality-Adjusted Time Without Symptoms and Toxicity (Q-TWiST)
Q-TWiST = TWiST + (uAE X AE) + (uPROG X PROG)
Utility
1.00
.75
.50
.25
0
TWiST
AE
PROG
DEATH
Time from Randomization

Figure 1. Health states for a Quality-adjusted Time Without Symptoms and Toxicity (Q-TWiST) analysis of zidovudine for asymptomatic HIV-infected patients. This illustrates the division of overall survival time into TWiST, AE (time after a severe adverse event), and Prog (time after HIV disease progression), and the weighting of these time periods using utility coefficients u_{AE} and u_{Prog}.

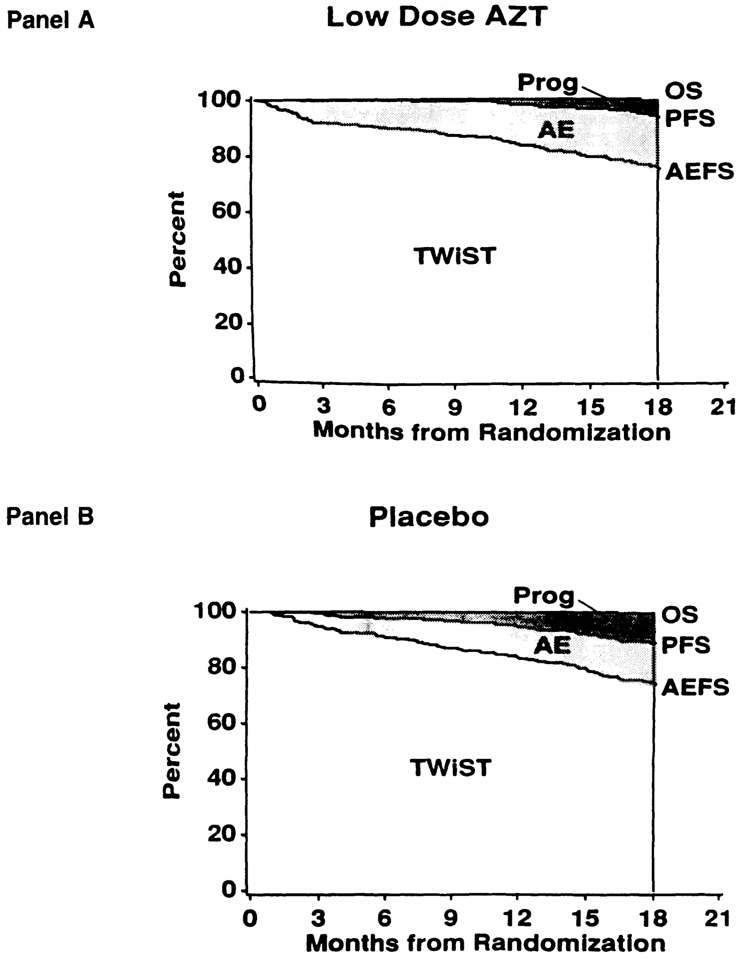


Figure 2. Partitioned survival plots for the 500-mg zidovudine group (panel A) and for the placebo group (panel B). In each graph, the area under the overall survival curve (OS) is partitioned by the Kaplan–Meier curves for HIV progression-free survival (PFS) and adverse event-free survival (AEFS). The areas between the Kaplan–Meier curves up to 18 months of follow-up give the restricted mean times spent in the health states TWiST, AE, and Prog.

Q-TWiST for all possible combinations of values of u_{AE} and u_{Prog} (ranging from 0.0 to 1.0; see figure 3). The solid line in figure 3 indicates the pairs of utility coefficients for which the Q-TWiST was equal for the two treatment groups. Pairs of utility coefficients above this line are those values of u_{AE} and u_{Prog} for which the Q-TWiST was greater for the zidovudine group, while those below the line favor placebo. The 95% confidence interval for the threshold utility line (upper boundary shown as the dashed line in figure 3) was calculated using the bootstrap method.

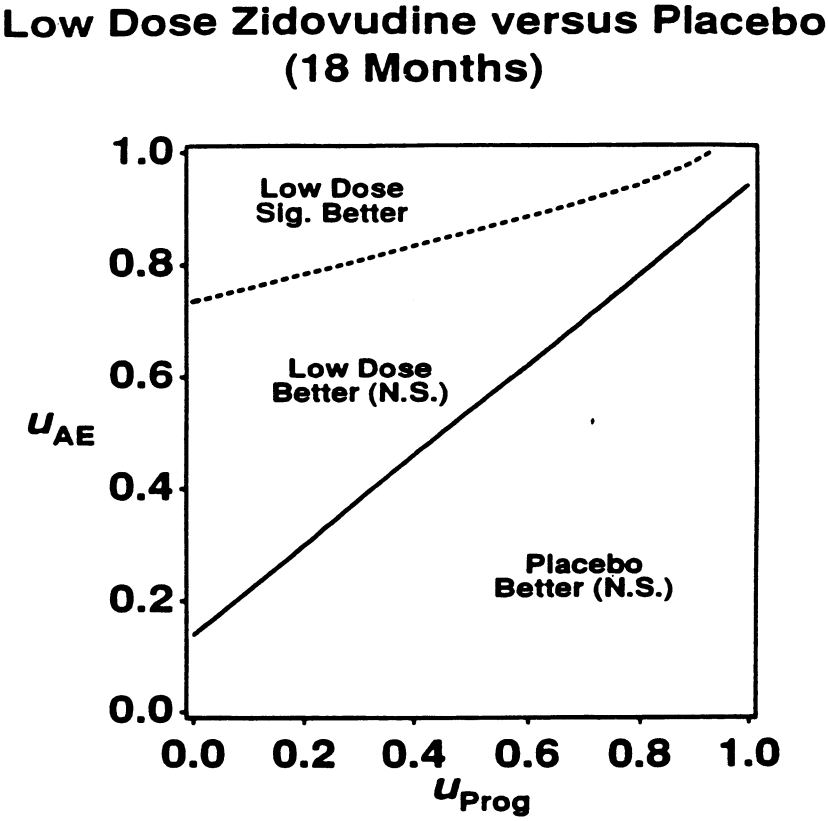


Figure 3. Threshold utility analysis comparing a 500-mg dose of zidovudine with placebo within 18 months of follow-up. The vertical axis shows the value of the time after an adverse event (u_{AE}), and the horizontal axis shows the value of the time after disease progression (u_{Prog}). The values for both range from 0.0 to 1.0, with a value of 1.0 indicating that the time is worth the same as TWiST and a value of 0.0 indicating that the time is worth nothing. The solid line is the threshold (based on values of u_{AE} and u_{Prog}) for which zidovudine and placebo groups have equal amounts of Q-TWiST. The dashed line shows the upper 95% confidence boundary for the threshold line (the lower confidence boundary is out of the possible range of utility values and hence is not shown). The lines divide the plot into three areas showing the utility values for which Q-TWiST was significantly longer in the zidovudine group (upper left), the values for which Q-TWiST was longer but not significantly so for the zidovudine group (middle region), and the values for which Q-TWiST was longer but not significantly so in the placebo group (lower right).

Subjective patient judgments provide the weights for the components of Q-TWiST influencing treatment comparisons. A traditional efficacy analysis, considering the delay of disease progression as the main endpoint, would assign a value of 1.0 to the coefficient u_{AE} and 0.0 to the coefficient u_{Prog}. For these utility values, the zidovudine yielded the better result. In fact, zidovudine provided significantly more Q-TWiST than placebo for all patients with utility coefficients in the upper left corner of figure 3, those with higher values for u_{AE} and lower values for u_{Prog}. The trade-off shifted away from zidovudine for patients who value time after adverse events as less than 1.0 and time after disease progression as greater than 0.0.

Extensions. The Q-TWiST method is very general, since it can be applied for any treatment comparison involving progressive health states that differ with respect to the patients' QOL. Current research is underway to apply a Q-TWiST analysis for treatment comparisons in rectal cancer, colon cancer, pediatric malignancies, cardiology, and neurologic diseases such as multiple sclerosis. In addition, techniques for eliciting utility coefficients directly from patients and incorporating these in the Q-TWiST analysis are being developed [65]. A Q-TWiST gain function has been introduced to display the treatment comparison over time [66]. The Q-TWiST method has been extended to incorporate prognostic factors using a Cox proportional hazards regression model [67]. Parametric estimates for the tails of the survival distributions have been used to project future treatment gains that might be achieved [68,69]. A version of Q-TWiST has also been developed for use in meta-analysis.

Conclusion

Although overall survival remains the most important single endpoint for evaluating treatment effectiveness, QOL evaluation plays an increasingly important role for therapeutic choice. The use of QOL assessments in clinical trials has suffered because these endpoints require special effort that is not yet a routine part of patient care or protocol conduct. The specific choice of QOL instruments is less important than a recognition and endorsement that QOL assessments in clinical trials represent an essential component of the research strategy.

Missing data, the censoring of QOL measurements by death, and the multiplicity of endpoints represent special methodological challenges in the analysis of QOL. Some of the methods used to handle these problems have been described. When designing a clinical trial, it is important to consider the type of QOL analysis that will be performed.

The Q-TWiST method was specifically developed to provide treatment comparisons within clinical trials that incorporate both quantity and quality of survival. These comparisons are based on a utility approach to weighting different health states according to patient preferences. Q-TWiST integrates the subjective aspects of QOL assessment into a survival-time analysis to provide a range of outcomes useful for patient care decision making.

Acknowledgments

We gratefully acknowledge the assistance of Dr. Kenneth Stanley, who provided valuable comments on this chapter. This chapter is dedicated to the memory of Philip A. Gelber, who died on March 9, 1994. Partial support was provided by Grant #PBR-53 from the American Cancer Society.

References

1. Maguire P (1986). Using measures of psychological impact of disease to inform clinical practice. In *Proceedings of the International Workshop on Quality of Life Assessment and Cancer Treatment*, V Ventafridda, FSAM van Dam, R Yancik, M Tamburini (eds.). Amsterdam: Excerpta Medica, 119–126.
2. Slevin ML (1992). Quality of life: Philosophical question or a clinical reality. *Br Med J* 305:466–469.
3. Weeks J (1992). Quality of life assessment: Performance status upstaged? *J Clin Oncol* 10:1827–1829.
4. Laupacis A, Sackett DL, Roberts RS (1988). An assessment of clinically useful measurements of the consequences of treatment. *N Engl J Med* 318(26):1728–1733.
5. Friedman LM, Furberg CD, Demets DL (1985). *Fundamentals of Clinical Trials*. Littleton, MA: PSG, 165.
6. Coates A, Gebski V, Signorini D, et al. (1992). Prognostic value of quality-of-life scores during chemotherapy for advanced breast cancer. *J Clin Oncol* 10:1833–1838.
7. Karnofsky DA, Abelmann WH, Craver LF, Burchenal JH (1948). The use of nitrogen mustards in the palliative treatment of carcinoma. *Cancer* 1:634.
8. Yates JW, Chalmer B, McKegney FP (1980). Evaluation of patients with advanced cancer using the Karnofsky Performance Status. *Cancer* 45:2220–2224.
9. Zubrod CG, Schneiderman M, Frei E, et al. (1960). Appraisal of methods for the study of chemotherapy in man. *J Chron Dis* 11:7.
10. Steinbroker O, Traeger CH, Battman RC (1949). Therapeutic criteria in rheumatoid arthritis. *JAMA* 140:659–662.
11. Criteria Committee of the New York Heart Association, Kossman CE, chairman (1964). *Diseases of the heart and blood vessels: Nomenclature and criteria for diagnosis*, 6th edition. Boston: Little, 112–113.
12. Patterson WB (1975). The quality of survival in response to treatment. *JAMA* 233(3):280–281.
13. Izsak FC, Medalie JH (1971). Comprehensive follow-up of carcinoma patients. *J Chron Dis* 24:179–191.
14. Burge PS, Prankard TAJ, Richards JDM, et al. (1975). Quality of survival in acute myeloid leukemia. *Lancet* 2:621–624.
15. Priestman TJ, Baum M (1976). Evaluation of quality of life in patients receiving treatment for advanced breast cancer. *Lancet* 1:899–900.
16. Stewart AL, Hays RD, Ware JE Jr. (1988). The MOS short-form general health survey: reliability and validity in a patient population. *Med Care* 26:724.
17. Spitzer WO, Dobson AJ, Hall J, et al. (1981). Measuring the quality of life of cancer patients. *J Chron Dis* 34:585–597.
18. Franshel S, Bush JW (1970). A health state index and its application to health-services outcomes. *Oper Res* 18(6):1021–1066.
19. Torrance GW, Thomas WH, Sackett DL (1972). A utility maximization model for evaluation of health care programs. *Health Service Res* 7:118.

20. Weinstein MC, Stason WB (1977). Foundations of cost-effective analysis for health and medical practices. *N Engl J Med* 296:716–721.
21. Levine MN, Guyatt GH, Gent M (1988). Quality of life in stage II breast cancer: an instrument for clinical trials. *J Clin Oncol* 6(12):1798–1810.
22. Cella DF, Tulsky DS, Gray G, et al. (1993). The functional assessment of cancer therapy scale: development and validation of the general measure. *J Clin Oncol* 11(3):570–579.
23. Weeks JC, Kohane I, Hayes D (1994). Health-related quality of life data collection in routine clinical care using hand-held computers. A protocol submitted at the Dana-Farber Cancer Institute.
24. Presser S, Schuman H (1980). The measurement of a middle position in attitude surveys. *Publ Opinion Q* 44:70–85.
25. Fayers PM, Jones DR (1983). Measuring and analyzing quality of life in cancer clinical trials: a review. *Stat Med* 2:429–446.
26. Kleinman A (1986). *Social Origins of Distress and Disorder; Depression, Neurasthenia, and Pain in Modern China*. New Haven: Yale University Press.
27. Guyatt GH, Feeny DH, Patrick DL (1993). Measuring health-related quality of life. *Ann Intern Med* 118(8):622–629.
28. Stewart AL (1990). Psychometric considerations in functional status instruments. In *Functional Status Measurement in Primary Care*, WONCA Classification Committee (eds.). New York: Springer-Verlag, 3–26
29. Deyo RA, Diehr P, Patrick DL (1991). Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. *Controlled Clin Trials* 12:142S–158S.
30. Cronbach LJ (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16:297–334.
31. Ganz PA, Schag CAC, Lee JJ, et al. (1992). The CARES: a generic measure of health-related quality of life for patients with cancer. *Qual Life Res* 1:19–29.
32. Aaronson NK, Bullinger M, Ahmedzai S (1988). A modular approach to quality-of-life assessment in cancer clinical trials. *Recent Results Cancer Res* 111:231–249.
33. Schipper H, Clinch J, McMurray A, et al. (1984). Measuring the quality of life of cancer patients: the functional living index–cancer: development and validation. *J Clin Oncol* 2:472–484.
34. Hürny C, Bernhard J, Gelber RD, et al., for the International Breast Cancer Study Group (1992). Quality of life measures for patients receiving adjuvant therapy for breast cancer. *Eur J Cancer* 28:118–124.
35. Kirk RE (1968). *Experimental Design: Procedures for the Behavioral Sciences*. Belmont, CA: Brooks/Cole.
36. Olschewski M, Schumacher M (1990). Statistical analysis of quality of life data in cancer clinical trials. *Stat Med* 9:749–763.
37. Schumacher M, Olschewski M, Schulgen G (1991). Assessment of quality of life in clinical trials. *Stat Med* 10:1915–1930.
38. Cox DR, Fitzpatrick R, Fletcher AE, et al. (1992). Quality-of-life assessment: can we keep it simple? *JR Stat Soc A* 155(3):353–393.
39. Nou E, Aberg T (1980). Quality of survival in patients with surgically treated bronchial carcinoma. *Thorax* 35:255–263.
40. Herson J (1980). Evaluation of toxicity: statistical considerations. *Cancer Treat Rep* 64:463–468.
41. Rosenman J, Choi NC (1982). Improved quality of life in patients with small cell carcinoma of the lung by elective radiation of the brain. *Int J Radiat Oncol* 8:1040–1043.
42. O'Brien PC (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* 40:1079–1087.
43. Pocock SJ, Geller NL, Tsiatis AA (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* 43:487–498.
44. Tandon PK (1990). Applications of global statistics in analyzing quality of life data. *Stat Med* 9:819–827.

45. Zwinderman AH (1990). The measurement of change of quality of life in clinical trials. *Stat Med* 9:931–942.
46. Gore S (1988). Integrated reporting of quality and length of life — A statistician's perspective. *Eur Heart J* 9:228–234.
47. Loewy JW, Kapadia AS, Hsi B, Davis BR (1992). Statistical methods that distinguish between attributes of assessment: prolongation of life versus quality of life. *Med Decision Making* 12:83–92.
48. Lagakos SW, Sommer CJ, Zelen M (1978). Semi-Markov models for partially censored data. *Biometrika* 65(2):311–317.
49. Silverstein MD, Albert DA, Hadler NM, Ropes MW (1988). Prognosis in SLE: comparison of Markov model to life table analysis. *J Clin Epidemiol* 41:623–633.
50. Levington A, Schulman J, Kammerman L, et al. (1980). A probability model of headache recurrence. *J Chron Dis* 53:407–412.
51. Myers LE, Paulson DE, Berry WR, et al. (1980). A time-dependent statistical model which relates current clinical status to prognosis: application to advanced prostatic cancer. *J Chron Dis* 33:491–499.
52. Kay R (1986). A Markov model for analyzing cancer markers and disease states in survival studies. *Biometrics* 42:855–865.
53. Pepe MS, Longton G, Thornquist M (1991). A qualifier for the survival function to describe the prevalence of a transient condition. *Stat Med* 10:413–421.
54. Goldhirsch A, Gelber RD, Simes RJ, et al., for Ludwig Breast Cancer Study Group (1989). Costs and benefits of adjuvant therapy in breast cancer: a quality-adjusted survival analysis. *J Clin Oncol* 7:36–44.
55. Feeny DH, Torrance GW (1989). Incorporating utility-based quality-of-life assessment measures in clinical trials. *Med Care* 27:S190–S204.
56. Gelber RD, Goldhirsch A (1986). A new endpoint for the assessment of adjuvant therapy in postmenopausal women with operable breast cancer. *J Clin Oncol* 4:1772–1779.
57. Gelber RD, Gelman RS, Goldhirsch A (1989). A quality of life oriented endpoint for treatment comparisons. *Biometrics* 45:781–795.
58. Glasziou PP, Simes RJ, Gelber RD (1990). Quality adjusted survival analysis. *Stat Med* 9:1259–1276.
59. Gelber RD, Goldhirsch A, Cavalli F, for the International Breast Cancer Study Group (1991). Quality-of-life-adjusted evaluation of adjuvant therapies for operable breast cancer. *Ann Intern Med* 114:621–28.
60. Gelber RD, Goldhirsch A, Hürny C, et al., for the International Breast Cancer Study Group (1992). Quality of life in clinical trials of adjuvant therapies. *J Natl Cancer Inst Monogr* 11:127–135.
61. Gelber RD, Lenderking WR, Cotton DJ, et al., for the AIDS Clinical Trials Group (1992). Quality-of-life evaluation in a clinical trial of zidovudine therapy in patients with mildly symptomatic HIV infection. *Ann Intern Med* 116:961–966.
62. Lenderking WR, Gelber RD, Cotton DJ, et al. (1994). Evaluation of the quality of life associated with zidovudine treatment in asymptomatic human immunodeficiency virus infection. *N Engl J Med* 330(11):738–743.
63. Kaplan EL, Meier P (1958). Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 54:457–481.
64. Volberding PA, Lagakos SW, Koch MA, et al. (1990). Zidovudine in asymptomatic human immunodeficiency virus infection: a controlled trial in persons with fewer than 500 CD-4 positive cells per cubic millimeter. *N Engl J Med* 322:941–949.
65. Weeks J, O'Leary J, Fairclough D, et al. (1994). The 'Q-tility index': a new tool for assessing health-related quality of life and utilities in clinical trials and clinical practice. *Proceedings of ASCO* 13:436.
66. Gelber RD, Goldhirsch A, Cole BF, for the International Breast Cancer Study Group, (1993) Evaluation of effectiveness: Q-TWiST. *Cancer Treat Rev* 19:73–84.
67. Cole BF, Gelber RD, Goldhirsch A (1993). Cox regression models for quality adjusted

survival analysis. *Stat Med* 12:975–987.

68. Gelber RD, Goldhirsch A, Cole BF, for the International Breast Cancer Study Group (1993). Parametric extrapolation of survival estimates with applications to quality of life evaluation of treatments. *Controlled Clin Trials* 14:485–489.
69. Cole BF, Gelber RD, Anderson KM, for the International Breast Cancer Study Group (1994). Parametric approaches to quality adjusted survival analysis. *Biometrics* 50:621–631.

Index