Tutorials in Biostatistics

Volume 1
Statistical Methods in Clinical Studies

Editor
R. B. D'Agostino

WILEY

Tutorials in Biostatistics

Tutorials in Biostatistics

Volume 1: Statistical Methods in Clinical Studies

Edited by

R. B. D'Agostino,
Boston University, USA

John Wiley & Sons, Ltd

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-470-02365-1

Typeset by Macmillan India Ltd
Printed and bound in Great Britain by Page Bros, Norwich
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Contents

2.2 Prognostic/Clinical Prediction Models

Part III CLINICAL TRIALS

3.1 Design

3.2 Monitoring

3.3 Analysis

Preface

The development and use of statistical methods has grown exponentially over the last two decades. Nowhere is this more evident than in their application to biostatistics and, in particular, to clinical medical research. To keep abreast with the rapid pace of development, the journal *Statistics in Medicine* alone is published 24 times a year. Here and in other journals, books and professional meetings, new theory and methods are constantly presented. However, the transitions of the new methods to actual use are not always as rapid. There are problems and obstacles. In such an applied interdisciplinary field as biostatistics, in which even the simplest study often involves teams of researchers with varying backgrounds and which can generate massive complicated data sets, new methods, no matter how powerful and robust, are of limited use unless they are clearly understood by practitioners, both clinical and biostatistical, and are available with well-documented computer software.

In response to these needs *Statistics in Medicine* initiated in 1996 the inclusion of tutorials in biostatistics. The main objective of these tutorials is to generate, in a timely manner, brief well-written articles on biostatistical methods; these should be complete enough so that the methods presented are accessible to a broad audience, with sufficient information given to enable readers to understand when the methods are appropriate, to evaluate applications and, most importantly, to use the methods in their own research.

At first tutorials were solicited from major methodologists. Later, both solicited and unsolicited articles were, and are still, developed and published. In all cases major researchers, methodologists and practitioners wrote and continue to write the tutorials. Authors are guided by four goals. The first is to develop an introduction suitable for a well-defined audience (the broader the audience the better). The second is to supply sufficient references to the literature so that the readers can go beyond the tutorial to find out more about the methods. The referenced literature is, however, not expected to constitute a major literature review. The third goal is to supply sufficient computer examples, including code and output, so that the reader can see what is needed to implement the methods. The final goal is to make sure the reader can judge applications of the methods and apply the methods. The tutorials have become extremely popular and heavily referenced, attesting to their usefulness. To further enhance their availability and usefulness, we have gathered a number of these tutorials and present them in this two-volume set.

Each volume has a brief preface introducing the reader to the aims and contents of the tutorials. Here we present an even briefer summary. We have arranged the tutorials by subject matter, starting in Volume 1 with 18 tutorials on statistical methods applicable to clinical studies, both observational studies and controlled clinical trials. Two tutorials discussing the computation of epidemiological rates such as prevalence, incidence and lifetime rates for cohort studies and capture–recapture settings begin the volume. Propensity score adjustment methods and agreement statistics such as the kappa statistic are dealt with in the next two tutorials. A series of tutorials on survival analysis methods applicable to observational study data are next. We then present five tutorials on the development of prognostics or clinical prediction models. Finally, there are six tutorials on clinical trials. These range from designing

and analysing dose response studies and Bayesian data monitoring to analysis of longitudinal data and generating simple summary statistics from longitudinal data. All these are in the context of clinical trials. In all tutorials, the readers is given guidance on the proper use of methods.

The subject-matter headings of Volume 1 are, we believe, appropriate to the methods. The tutorials are, however, often broader. For example, the tutorials on the kappa statistics and survival analysis are useful not only for observational studies, but also for controlled clinical studies. The reader will, we believe, quickly see the breadth of the methods.

Volume 2 contains 16 tutorials devoted to the analysis of complex medical data. First, we present tutorials relevant to single data sets. Seven tutorials give extensive introductions to and discussions of generalized estimating equations, hierarchical modelling and mixed modelling. A tutorial on likelihood methods closes the discussion of single data sets. Next, two extensive tutorials cover the concepts of meta-analysis, ranging from the simplest conception of a fixed effects model to random effects models, Bayesian modelling and highly involved models involving multivariate regression and meta-regression. Genetic data methods are covered in the next three tutorials. Statisticians must become familiar with the issues and methods relevant to genetics. These tutorials offer a good starting point. The next two tutorials deal with the major task of data reduction for functional magnetic resonance imaging data and disease mapping data, covering the complex data methods required by multivariate data. Complex and thorough statistical analyses are of no use if researchers cannot present results in a meaningful and usable form to audiences beyond those who understand statistical methods and complexities. Reader should find the methods for presenting such results discussed in the final tutorial simple to understand.

Before closing this preface to the two volumes we must state a disclaimer. Not all the tutorials that are in these two volumes appeared as tutorials. Three were regular articles. These are in the spirit of tutorials and fit well within the theme of the volumes.

We hope that readers enjoy the tutorials and find them beneficial and useful.

RALPH B. D'AGOSTINO, SR. EDITOR
Boston University
Harvard Clinical Research Institute

Preface to Volume 1

This first volume of *Tutorials in Biostatistics* is devoted to statistical methods in clinical research. By this we mean statistical methods applied to medical problems involving human beings, either as members of populations or groups in observational and epidemiological research or as participants in clinical trials. The tutorials are divided into three parts. Here we briefly mention the general themes of each part and the articles within them.

Part I is on observational studies and epidemiology. These articles clarify the uniqueness and complications that arise from observational data and present methods to obtain meaningful and unbiased inferences. Section 1.1 is devoted to epidemiology and contains two tutorials. The first, by Beiser, D'Agostino, Seshadri, Sullivan and Wolf, presents a thorough discussion of epidemiological event rates such as incidence rates and lifetime risks, clarifies issues such as competing risks in the calculation of these rates and includes computer programs to carry out computations. The second tutorial, by Chao, Tsay, Lin, Shau and Chao, describes the computation of epidemiological rates for capture-recapture data such as would be obtained from multiple surveys attempting to estimate the disease prevalence rate for, say, hepatitis A virus or diabetes in a population. The issue of minimizing biases is discussed. Section 1.2, on adjustment methods, contains one article by Ralph D'Agostino, Jr., on the use of propensity scores for reducing bias in treatment comparisons from observational studies. This tutorial has become a standard reference for propensity scoring. The article from Section 1.3, on agreement statistics, by Kraemer, Periyakoil and Noda, covers in detail the use of the kappa statistic in medical research.

Section 1.4 presents three tutorials devoted to survival methods applicable to observational studies. First, Bull and Spiegelhalter clearly identify the complications and other issues involved in using survival methods in observational studies (in contrast to clinical trials). Interval estimation and binary outcomes in longitudinal studies are then developed in the next two tutorials by Lindsey and Ryan and by Carlin, Wolfe, Coffey and Patton. The latter two tutorials have uses beyond survival analysis in observational studies. We group them in this part of the volume mainly for convenience. The reader should quickly see the broader applicability of the methods and not be limited by our classification.

Part II is concerned with prognostic/clinical prediction models and contains two sections. Here the aim is to present methods for developing mathematical models that can be used to identify people at risk for an outcome such as the development of heart disease or for the prognosis of subjects with certain clinical characteristics such as cancer tumour size. Some of these tutorials have become major references for clinical prediction model development. Section 2.1 contains one article by Mazumdar and Glassman on categorizing prognostic variables. The question is often how best to dichotomize a diagnostic variable so that it can be used in a clinical setting. Issues such as multiple testing often render useless such 'obvious' methods as trying to find the best cut point. A careful review of the field is presented and helpful suggestions abound.

Section 2.2, 'Prognostic/Clinical Prediction Models', presents in four detailed tutorials methods for developing and evaluating multivariable clinical prediction models. The first, by D'Agostino, Belanger, Markson, Kelly-Hayes and Wolf, illustrates how to deal with a large set of potentially useful prediction variables. Methods such as principal components analysis and hierarchical variable selection methods for survival analysis are highlighted. The next two articles have Frank Harrell as the first author and deal in detail with developing prediction models for time to event, binary and ordinal outcomes. (The first is authored by Harrel, Lee and Mark, and the second by Harrell, Margolis, Gove, Mason, Mulholland, Lehmann, Muhe, Gatchalian and Eichenwald.) Questions of model development are explored completely, as are issues of making predictions and concerns about validation. The last tutorial in Section 2.2 deals with estimating prognosis based on observational data such as are obtainable in a registry. It is authored by DeLong, Nelson, Wong, Pryor, Peterson, Lee, Mark, Califf and Pauker. These four tutorials are among the best literature sources for the development and appropriate use of clinical prediction models.

Part III is on clinical trials and contains three sections. While given as tutorials, the articles of this section are innovative in understanding as well as in the presentation of the issues and methods. Section 3.1 contains a clever article by Wong and Lachenbruch on the optimal design of dose response studies. Section 3.2, on monitoring in clinical trails, contains an article on Bayesian data monitoring by Fayers, Ashby and Parmar pointing to the benefits of a Bayesian analysis even in this setting. Section 3.3 contains four articles on analysis. These bring together ideas and methods available for use, but not presented elsewhere with such completeness and clear focus. They fill a serious void and add wonderfully to the field. The first, by Albert, deals with longitudinal clinical trial data analysis. The second, by Senn, Stevens and Chaturvedi, deals with generating simple summary numbers from repeated measures studies so that the analysis and interpretations of the study are intuitive and meaningful. The next article, by Agresti and Hartzel, concerns binary data from multi-centre trials. Lastly, Chuang-Stein and Agresti discuss dose responses with ordinal data.

We hope these 18 tutorials will be of use to readers.

Part I
OBSERVATIONAL STUDIES/ EPIDEMIOLOGY

1.1 Epidemiology

Computing estimates of incidence, including lifetime risk: Alzheimer's disease in the Framingham Study. The Practical Incidence Estimators (PIE) macro

Alexa Beiser[1,*,†], Ralph B. D'Agostino, Sr[2], Sudha Seshadri[3], Lisa M. Sullivan[1] and Philip A. Wolf[3]

[1] *Department of Epidemiology and Biostatistics, Boston University School of Public Health, Boston, MA, U.S.A.*
[2] *Department of Mathematics, Boston University, Boston, MA, U.S.A.*
[3] *Department of Neurology, Boston University School of Medicine, Boston, MA, U.S.A.*

SUMMARY

The incidence of disease is estimated in medical and public health applications using various different techniques presented in the statistical and epidemiologic literature. Many of these methods have not yet made their way to popular statistical software packages and their application requires custom programming. We present a macro written in the SAS macro language that produces several estimates of disease incidence for use in the analysis of prospective cohort data. The development of the Practical Incidence Estimators (PIE) Macro was motivated by research in Alzheimer's Disease (AD) in the Framingham Study in which the development of AD has been prospectively assessed over an observation period of 24 years. The PIE Macro produces crude and age-specific incidence rates, overall and stratified by the levels of a grouping variable. In addition, it produces age-adjusted rates using direct standardization to the combined group. The user specifies the width of the age groups and the number of levels of the grouping variable. The PIE macro produces estimates of future risk for user-defined time periods and the remaining *lifetime risk* conditional on survival event-free to user-specified ages. This allows the user to investigate the impact of increasing age on the estimate of remaining lifetime risk of disease. In each case, the macro provides estimates based on traditional unadjusted cumulative incidence, and on cumulative incidence adjusted for the competing risk of death. These estimates and their respective standard errors, are provided in table form and in an output data set for graphing. The macro is designed for use with survival age as the time variable, and with age at entry into the study as the left-truncation variable; however, calendar time can be stubstituted for the survival time variable and the left-truncation variable can simply be set to zero. We illustrate the use of the PIE macro using Alzheimer's Disease incidence data collected in the Framingham Study.

* Correspondence to: Alexa Beiser, Department of Epidemiology and Biostatistics, Boston University School of Public Health, 715 Albany Street, Boston, MA 02118, U.S.A.
† E-mail: alexab@bu.edu

Contract/grant sponsor: Framingham Heart Study of the National Heart, Lung and Blood Institute
Contract/grant sponsor: National Institute of Aging; contract/grant number: 5-R01-A608122-11
Contract/grant sponsor: NIH/NHLBI; contract/grant number: N01-HC-38038

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
© 2004 John Wiley & Sons, Ltd. ISBN: 0-470-02365-1

1. INTRODUCTION

The Framingham Heart Study is a population-based cohort study wherein subjects have been evaluated biennially for cardiovascular risk factors and cardiovascular endpoints since 1948 [1]. Over the years, additional data have been accumulated on sociodemographic and life-style factors and the cohort members have been screened for a variety of novel risk factors and non-cardiac disease conditions such as dementia, osteoporosis, cancer, visual loss and hearing impairment. Of the original 5209 subjects, 920 subjects are alive; their current mean age is 86 years. Thus we have a well-characterized cohort of 'old-old' subjects, traditionally defined as subjects over the age of 80 or 85 years. This is the fastest growing segment of the U.S. population and the study of the incidence of dementia and risk factors associated with dementia in this population is of enormous public health importance.

The incidence of disease is estimated in medical and public health applications using a variety of different techniques. Most of these techniques are discussed in detail in books on survival analysis [2–7], epidemiologic methods [8, 9] or biostatistical methods [10, 11]. Other techniques have been presented in the statistical or epidemiologic literature [12–17], or have simply been applied in the medical or public health literature [18]. Many of these methods have not yet made their way to popular statistical software packages and their application requires custom programming.

We present a macro written in the SAS macro language that produces several estimates of disease incidence for use in the analysis of prospective cohort data. This work was motivated by research in Alzheimer's disease (AD) in the Framingham Study in which the development of AD has been prospectively assessed over an observation period of 24 years. Our goal is to use these data to estimate: (i) crude and age group-specific yearly incidence of AD; (ii) age-adjusted yearly incidence of AD within selected subgroups; and (iii) the future risk of developing AD conditional on survival dementia-free to selected ages. We estimate future risk for predefined periods and the remaining lifetime risk, using traditional unadjusted cumulative incidence (UCI), and cumulative incidence adjusted for the competing risk of death (ACI).

2. MOTIVATION

The estimation of yearly incidence is relatively straightforward; however, in prospective studies such as the Framingham Study, there are several issues that make the estimation of cumulative incidence difficult. In order to generate a valid estimate of future risk, including the lifetime risk of developing AD, we must address the following. First, we must consider that individuals are followed for different periods of time. Second, the time origin must be defined such that individuals are comparable at the time origin. Third, we must account for subjects entering the observation period at different ages. Finally, we must address the impact of the competing risk of death. We discuss each issue below.

2.1. Individuals are followed for different periods of time

If we could follow every subject in the sample until either the end of the study or until they developed AD, we could directly estimate the probability of developing AD as the proportion of subjects in our sample who developed AD during the observation period, or the *cumulative*

incidence [10]. As is generally the case in a long-term prospective study, there are many subjects who are not observed for the entire observation period and we must use survival analysis techniques to estimate the cumulative incidence. The primary technique we will rely on is a modified Kaplan–Meier method.

2.2. Time origin

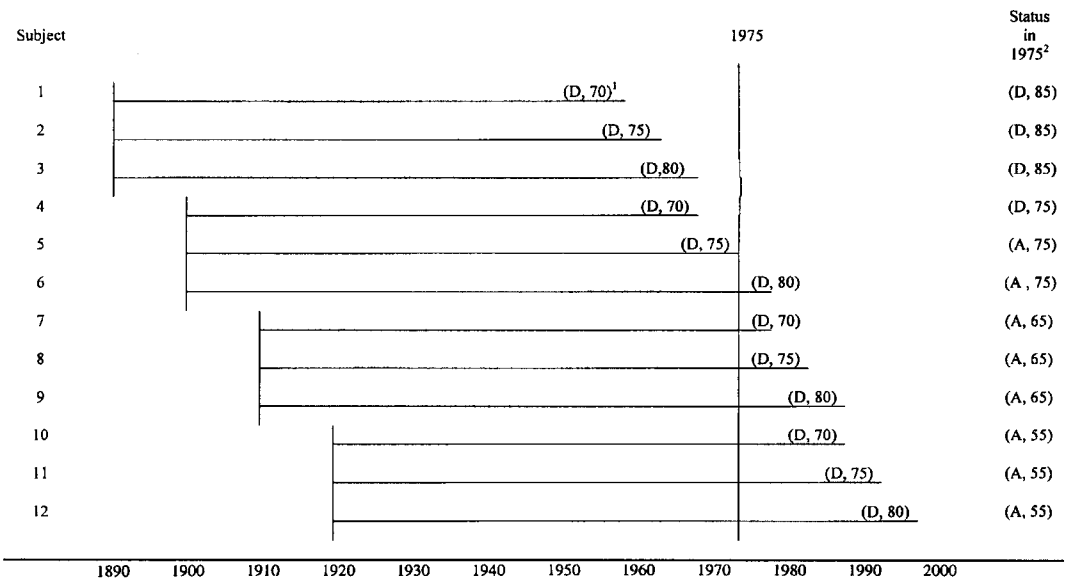
Prospective studies are often analysed using survival methods in which the dependent variable is survival time defined as the time from a designated origin to the event of interest (for example, diagnosis of AD). The time origin can be defined in several ways including the date of entry into a study or date of birth. If the time origin is defined as the date of entry into a study, an individual's time to event coincides with his/her time-on-study. If the time origin is defined as birth date, an individual's time to event is his/her survival age. However the time origin is defined, it is critical that individuals are comparable at the time origin [19]. In our case, subjects free of AD in 1975 enter the observation period at different ages. One subject may enter the observation period at age 50 years while another may enter at 70 years of age. Because the risk of AD is known to increase with age, it is important to take into account the different ages at which subjects enter the study. In this case, if the time origin is 1975, accounting for age may be done by adjusting for age as a covariate or by stratifying by age at entry (supposing we are able to categorize ages in a manageable way). This type of adjustment may not be sufficient.

Korn *et al.* [20] argue that survival age is more appropriate as a time scale than time-on-study for most outcomes. This approach assumes that the risk of development of the event of interest is more likely to change as a function of age than as a function of calendar time. This is certainly true for the development of dementia.

2.3. Subjects enter the observation period at different ages: selected risk set strategy

The survival age approach is ideal for population studies in which individuals are studied from their birth to event onset (for example, AD) or until they are right-censored with no event. Unfortunately, full survival information on an entire population is usually not available; rather, survival information is collected prospectively on a sample of event-free individuals selected from the population at the beginning of an investigation. When the time origin is defined to be birth, an individual's follow-up period is the sum of the observation period and the (event-free) period from birth to entry into the study. Individuals with incident events before the observation period are excluded from the entire study as is their follow-up time from birth to event. Thus follow-up time that occurred before the study is included, but *only if it is event-free.*

Very few, if any, subjects develop AD before the age of 65. In fact, in our data, the earliest AD case was diagnosed at age 69, and we only consider cases diagnosed at ages 70 or older. Consider a subject who enters the observation period at age 45 in 1975 and is observed for 24 years to 1998. In 1998 the subject reaches age 69 and is free of AD. In our estimate of the incidence rate, this subject contributes 24 person years (24 years free of AD) to the denominator. Including these person-years in the denominator decreases the estimate of incidence. A more useful estimator would be based on a restricted set of subjects, subjects who are truly 'at risk of developing AD' during the observation period (for example, subjects who are at least 70 years of age during the observation period).



1: Event status (D=dead, A=alive) and survival age
2: Event status (D=dead, A=alive) and age in 1975 (at start of observation period)

Figure 1. Lifetimes of 12 hypothetical patients.

Analytic approaches that rely on person-years at risk can underestimate hazards and incidence rates. For example, the hazard at age 70 years is the conditional probability of event onset at age 70, or the number of events at age 70 divided by the number of individuals at risk at age 70. Individuals who contribute time that occurred before the observation period (in other words, individuals who were older than 70 when the study began) appear in the denominator, but not in the numerator.

The bias in estimating hazards and incidences can be removed by excluding follow-up time that occurs before the observation period, by *left-truncation.* The risk set, at any age, must include only those individuals who were at risk at that age during the observation period.

Consider, for example, a hypothetical population of 12 individuals whose lifetimes are displayed in Figure 1. Suppose the event of interest is death. The population hazard at age 70 is defined as the ratio of the number of deaths at age 70 to the number of subjects in the risk set (defined as the number of subjects at risk entering age 70) and is equal to $4/12 = 0.33$. If an investigation is initiated in 1975 (see Figure 1), then 8 subjects are included in the investigation (subjects 5–12 who are alive at the beginning of the observation period in 1975). In a standard survival analysis of this 1975 cohort using survival age as the dependent variable, each of the 8 subjects contributes the number of years from his/her birth to survival age. The risk set entering age 70 includes 8 subjects, of whom 2 died at age 70 (Subjects 7 and 10). The population hazard is underestimated as $2/8 = 0.25$. Subjects 5 and 6 are older than 70 years of age at the beginning of the observation period (1975); they were both at risk at age 70, however, they were not at risk at

age 70 during the observation period. These subjects should be excluded from the risk set at age 70 (in 1975). If we restrict the risk set at age 70 to only those who are at risk at age 70 during the observation period, the risk set includes 6 subjects, and we correctly estimate the population hazard as $2/6 = 0.33$. This approach reflects a selected risk set strategy and we use this strategy in our computations.

2.4. Competing risks

The estimation of cumulative incidence of AD is complicated by a fairly common situation: the development of AD is subject to the competing risk of death. Subjects who die during the observation period are treated as censored observations in traditional survival analytic techniques such as the Kaplan–Meier method [21]. This method is inappropriate as it assumes that failure from the event of interest is still possible beyond the time at which the censoring occurred. For example, a person who dies of cardiovascular disease cannot develop AD and should not contribute to the estimate of development of AD. Gooley [13] shows that the potential contribution of censored observations to the probability of failure from the event of interest is distributed among those subjects remaining at risk. However, the potential contribution of a subject who has died should be zero. Treating such subjects as censored inflates the estimate of cumulative incidence. Various analytic solutions to the problem of competing risks have been proposed and implemented [12–17], but there is still no software available that addresses this issue.

We provide estimates of both the unadjusted cumulative incidence (UCI) and the cumulative incidence adjusted for the competing risk of death (ACI). (Note that these are generally referred to in the literature as 1-KM – the complement of the Kaplan–Meier estimate of survival – versus CI [12, 13]). The ACI is useful as an estimate of the probability of actually developing AD, while the UCI estimates the probability of developing AD assuming no competing risk (that is, all subjects living for the entire lifespan). The former estimator is particularly useful from a public health standpoint as it allows the estimation of the numbers of cases of AD one can expect in a given population. Further, by adjusting the observed cumulative incidence using the mortality experience of a 'standard population' one can estimate a standardized lifetime risk. In some diseases which appear to be associated with aging *per se*, the exponential rise in annual incidence with increasing age is balanced by the exponential decrease in life expectancy seen with age, resulting in a relatively invariant estimate of the lifetime risk in elderly individuals. Thus for instance, in the Framingham Study, the lifetime risks of Alzheimer's disease [18] and congestive heart failure [22] were found to remain relatively constant with increasing age beyond 65 years. The ability to generate a single sex-specific estimate of lifetime risk is useful in educating the public regarding the true risk of the disease. The unadjusted cumulative incidence may be useful in the pathophysiological investigation of potential risk factors for AD. As an example, cigarette smoking may appear to provide protection from AD when the ACI is used to estimate cumulative incidence. This could be a simple consequence of the fact that cigarette smoking is associated with increased mortality, thus decreasing the observed incidence of AD. Smoking may, however, increase the physiological risk of AD; this would be seen only if the UCI is used to estimate cumulative incidence (Seshadri *et al.*, submitted to the American Academy of Neurology, 2000).

2.5. Statistical software

Many standard statistical computing packages do not handle these issues in a straightforward manner, if at all. Even the calculation of one-year incidence rates and age-adjusted rates using

direct standardization requires a certain amount of programming. Many standard statistical software packages do not exclude follow-up that occurred before the observation period and thus underestimate hazards and cumulative incidence. For example, SAS *Proc Lifetest* provides estimates of cumulative incidence using the Kaplan–Meier method, but has no mechanism for left-truncation. SAS *Proc Phreg* performs proportional hazards modelling and allows left-truncation but does not consider the adjustment for competing risk that is often necessary. We developed an SAS macro that produces: one-year incidence rates by age group; age-adjusted rates to compare rates among the levels of a grouping variable; estimates of traditional Kaplan–Meier cumulative incidence; and estimates of cumulative incidence adjusted for the competing risk of another event. Confidence intervals for each of the cumulative incidence estimates can also be provided.

3. THE FRAMINGHAM STUDY

The Framingham Study is a longitudinal study of 5209 participants (2336 men and 2873 women) which began in 1948 [1]. Participants have been examined in biennial exam cycles from 1948 to the present. At study onset, the initial ages ranged from 28 to 62 years.

3.1. Neuropsychological assessment in the Framingham Study

Several standardized neuropsychological batteries have been administered to participants at biennial exam cycles beginning with exam 14 in 1975/1976 to prospectively assess dementia. The Kaplan Albert (KA) battery [23] was introduced in exam 14 and includes sub-tests taken or derived from: (i) the original Weschler Memory Scale (including the logical memory, logical memory-delayed, logical-memory retained tests); (ii) sub-tests of the Weschler Adult Intelligence Scale (including similarities, digit span forward and digit span backward tests); and (iii) a measure of word fluency taken from the Aphasia Examination. Starting with exam 17 in 1982/1983, the Folstein Mini Mental State Examination (MMSE) [24] has been administered to participants on a biennial basis.

3.2. Generating the Framingham Dementia Cohort

The Framingham Dementia Cohort includes $n = 2611$ participants who were dementia-free in 1975. The criteria outlined below used to determine inclusion in the dementia cohort are strict enough to ensure that members were indeed dementia-free. Participants had to pass the Kaplan–Albert battery or the MMSE (score at least 24 of a possible 30 points) to be included. The dementia cohort contains people who satisfied the following (see Figure 2):

(i) Passed the Kaplan–Albert battery in 1975 at exam 14 ($n = 2083$), or did not take the Kaplan–Albert Battery at exam 14 (as it was introduced part way through the cycle) and either:
(ii) passed the MMSE at exam 17 ($n = 474$), or
(iii) did not take the MMSE at exam 17 but passed it at exam 18 ($n = 38$), or
(iv) did not take the MMSE at exams 17 or 18 but passed it at exam 19 ($n = 17$).

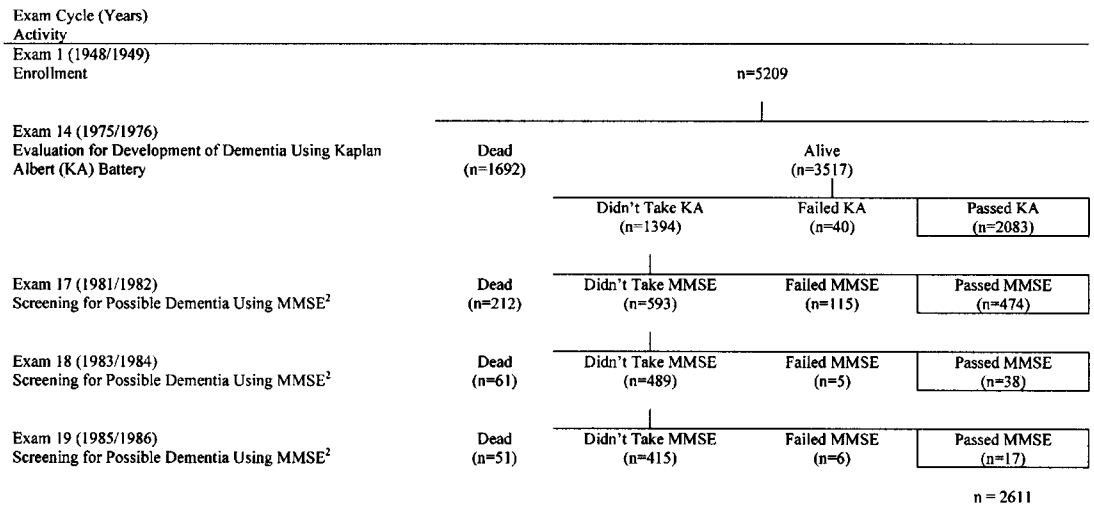


Figure 2. The Framingham dementia cohort: participants dementia free in 1975 ($n = 2611$). (Cases in the Framingham dementia cohort satisfied the following: passed the KA exam ($n = 2083$), or did not take the KA exam and either passed the MMSE at exam 17 ($n = 474$), or did not take the MMSE at exam 17 but passed the MMSE at exam 18 ($n = 38$), or did not take the MMSE at exams 17 or 18 but passed at exam 19 ($n = 17$). MMSE is Mini Mental State Examination.)

3.3. Dementia surveillance protocol; Identification of cases

Framingham Study participants are examined for cognitive decline according to a standardized protocol. At each biennial examination, subjects can be flagged for further evaluation based on: (i) self-report or family report of memory loss; (ii) referral by the physician conducting the biennial examination; or (iii) their performance on the MMSE. For this purpose, poor performance on the MMSE is defined as an absolute score below an education-based cut-off, a score more than 3 points lower than the score at the previous assessment or a score more than 5 points less than any previous score. Subjects may also be referred for evaluation by their primary care physician or by another source, such as the ongoing study 'Precursors of Stroke Incidence and Prevalence' in the same cohort.

A neuropsychologist reviews participants flagged by the initial screen, performs additional neuropsychological assessments and enters the participant in the dementia tracking protocol. Simultaneously but independently, a neurologist evaluates the flagged participants and, based on the neurology examination, classifies patients as not demented, or as mildly, moderately or severely demented. Patients classified as not demented or as mildly demented continue to be monitored with the dementia tracking protocol that includes annual neuropsychological re-examination and neurological re-evaluation as indicated, usually at least once every two years. Those who are classified as moderately or severely demented are evaluated at an in-depth dementia review.

A panel of at least two neurologists and one neuropsychologist conducts the dementia review. The review is based on neurology examination findings, neuropsychological assessments,

Framingham Study records, hospital and nursing home records, brain imaging, information from primary care physicians, and data gathered by telephone interview of family or next of kin. The panel is responsible for determining the presence of definite dementia (based on DSM-III [25] and, later, DSM-IV [26] criteria). At Framingham, a diagnosis of definite dementia also requires the presence of symptoms for at least 6 months and presence of 'moderate' dementia (severity not less than 1 on the Clinical Dementia Rating Scale). Participants who do not satisfy the criteria for definite dementia continue to be monitored according to the neuropsychology tracking protocol and are rereviewed by the dementia review panel as necessary. Participants identified as definitely demented are assigned a subtype of dementia, a year of dementia onset, and a year of diagnosis, that is, the year in which the criteria for a diagnosis of dementia were first fully satisfied. Criteria established by the National Institute of Neurological and Communicative Disorders – Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA [27]) are used in the diagnosis of probable Alzheimer's disease.

Participants who have suffered a stroke (usually after the onset of dementia) but do not meet the DSM-IV criteria for vascular dementia are classified as 'dementia complicated by stroke, relationship unknown'. These participants are combined with those classified as probable AD to estimate the upper limit of the risk of AD in the Framingham population.

Members of the Framingham Study cohort who have not participated in recent biennial examinations may be identified for dementia review after their death when a separate panel reviews their medical history as part of the Framingham Study protocol.

3.4. Dementia cohort follow-up status

Participants included in the dementia cohort ($n = 2611$) were followed from 1975 to 1998. Each participant is classified as a case (that is, he/she satisfied the criteria for dementia as outlined in Section 3.3 during the follow-up period), a death (that is, he/she died dementia-free during the follow-up period) or as a censored observation. Each classification is described in detail below. In all cases, follow-up times (that is, until development of dementia, death or censoring) are based on the last year of follow-up as opposed to an exact date.

The last year of follow-up for participants classified as cases is the year of diagnosis of dementia. This is necessarily later than the year of dementia onset, but may be determined using objective criteria as compared to the year of dementia onset.

Participants classified as deaths died during the follow-up period and were determined to be dementia-free by (i) post-mortem dementia review or (ii) post-mortem medical record review confirming the absence of cognitive decline. The last year of follow-up for this group is their year of death.

All other members of the dementia cohort are classified as censored. This group contains participants who are dead and (i) have not yet had a post-mortem medical record review, (ii) were referred for dementia review after the post-mortem medical record review but have not yet been reviewed by the dementia review panel, or (iii) were not identified as demented at initial review by the dementia review panel but need a post-mortem re-review. In addition, this group contains participants who are alive and (i) have never have been flagged for cognitive evaluation, or (ii) were flagged for cognitive evaluation and are being monitored according to the neuropsychology tracking protocol. The participants in this group can be in any stage of the dementia surveillance protocol but are all censored in the last year they could be classified as not demented. This is the year of their last normal neurology exam (in which they were classified as not demented or as

mildly demented) or the year of their last normal MMSE (as defined using the education-based cut-off), whichever is later.

4. METHODS

In this section we establish notation and provide the formulae that are operationalized in the macro. The description of the SAS macro follows in Section 5.

4.1. Data structure

Alzheimer's disease develops insidiously over time as compared to other events, such as stroke, that occur suddenly. For this reason, its development is reported using a year (rather than a specific date) of diagnosis. To be consistent, we use years as the measure of time for all analyses. Consider the following three subjects:

Subject	Date of birth	Year of AD diagnosis	Date of death
A	07/03/1903	1980	XXX
B	01/23/1905		11/29/1978
C	03/03/1905	1975	06/22/1989

The traditional data structure for a life-table analysis of these data would consider each subject's contribution in terms of years on study (starting in 1975). The risk set would initially contain all subjects and its size would be a non-decreasing function of time. For example, subject A is at risk of developing AD at entry in 1975 and entering each subsequent year to 1980. Subject A thus contributes six years of follow-up, with an event in the sixth year. Subject B contributes four years and is censored in the fourth year in 1978. Subject C contributes one year (1975) with an event in that year. The traditional data structure for these three subjects is as follows:

Years on study	Number of subjects entering risk set
1	3
2	2
3	2
4	2
5	1
6	1
7	0

We can reorganize the data with age as the time scale and also allow for the left-truncation necessary to account for subjects entering at different ages. Each member of the dementia cohort provides one observation with the following variables:

(i) *entry age*, age, in years, at entry in the observation period (that is, age in 1975);
(ii) *survival age*, age, in years, in their last year of follow-up;

(iii) *event status*, an indicator variable coded 1 for subjects who developed AD during the observation period and 0 for those who did not;
(iv) *any event status*, an indicator variable coded 1 for subjects who either developed AD or died during the observation period and 0 for those who did neither.

For example, subject A has *entry age* = 72 (1975–1903); *survival age* = 77 (1980–1903); *event status* = 1; and *any event status* = 1. This subject contributes 6 years of data at ages 72, 73, 74, 75, 76 and 77. Notice that subject A is considered to contribute an entire year of age 72 follow-up (in 1975) although this subject will not actually be 72 until 3 July 1975. We also consider that subject A becomes demented at age 77 (in 1980) when the age at diagnosis could really be as young as $76\frac{1}{2}$. It is not possible to more accurately assign a time of diagnosis and we choose calendar year to index subjects' ages.

The second subject, subject B, has *entry age* = 70 (1975–1905); *survival age* = 73 (1978–1905); *event status* = 0; *any event status* = 1, and contributes 4 years, at ages 70, 71, 72 and 73. Subject C has *entry age* = 70 (1975–1905); *survival age* = 70 (1975–1905); *event status* = 1; and *any event status* = 1. This subject contributes one year at age 70.

We define the following notation to summarize the data structure we will use to perform analyses with age as the time scale:

r_A = the number of persons at risk entering age A;
e_A = the number of incident events of interest (for example cases of AD) at age A;
w_A = the weighted number of persons at risk entering age A, computed by assigning a weight of 0.5 observations censored free of the event of interest during age A and a weight of 1 to all other observations (the actuarial method);
c_A = the number of persons who fail due to either the event of interest, or due to the competing risk (here, death).

In this case the size of the risk set is not a non-decreasing function of time; rather it decreases as subjects fail or are censored, but increases as subjects age into the study period. A summary of data on the three subjects in the example above is as follows:

Age	r_A	e_A	w_A	c_A
70	2	1	2	1
71	1	0	1	0
72	2	0	2	0
73	2	0	1.5	1
74	1	0	1	0
75	1	0	1	0
76	1	0	1	0
77	1	1	1	1

4.2. One-year incidence rates by age group

We summarize the age-specific data described in Section 4.1 by collapsing the age-specific data into age groups. The one-year incidence rate (per 1000 person-years) in each age group G is

calculated as the total number of incident events divided by the total number of (weighted) person-years at risk, times 1000:

$$\mathrm{IR}_G = 1000 \times \{(\Sigma_{A \in G}(e_A))/(\Sigma_{A \in G}(w_A))\} \tag{1}$$

where the summations are over the ages, A, included in the age group. For example, the 1-year incidence rate per 1000 person-years for the age group 70–74 is

$$\mathrm{IR}_{70\text{–}74} = 1000 \times \{(1)/(2 + 1 + 2 + 1.5 + 1)\} = 1000(1/7.5) = 133.33.$$

The crude 1-year incidence rate (over all ages) is estimated as the total number of incident events divided by the total number of person-years at risk times 1000:

$$\mathrm{IR}_C = 1000 \times \{(\Sigma_{\mathrm{all}\, A}(e_A))/(\Sigma_{\mathrm{all}\, A}(w_A))\} \tag{2}$$

where the summations are over all ages.

4.3. Age-adjusted rates

We use direct standardization to calculate age-adjusted rates for comparison of rates among levels of a grouping variable. The combined group (over all levels of the grouping variable) is used as the standard population. For example, if the grouping variable has two levels and we define, in age group G, the 1-year incidence rates in levels one and two, IR_{1G} and IR_{2G}, respectively, then the age-adjusted rates are

$$\mathrm{IR}_{1A} = \Sigma_{\mathrm{all}\, G}(\mathrm{IR}_{1G})(p_G)$$

and

$$\mathrm{IR}_{2A} = \Sigma_{\mathrm{all}\, G}(\mathrm{IR}_{2G})\,(p_G) \tag{3}$$

where $p_G = \{w_G/(\Sigma_{\mathrm{all}\, G}\,(w_A)\}$ is the proportion of (weighted) person-years in age group G.

4.4. Kaplan–Meier estimate of unadjusted cumulative incidence (UCI)

Suppose that $t_1 < t_2 < \ldots < t_j < \ldots < t_J$ are the ordered failure times among N subjects ($J < N$), e_j is the number of patients who fail from the event of interest (AD) at time t_j, and r_j is the number of patients at risk at time t_j (their failure or censoring times are greater than or equal to t_j). The Kaplan-Meier estimate of survival beyond time t (for example, probability of not developing AD) is given by

$$\hat{S}(t) = \Pr\{T > t\} = \prod_{j=1}^{k}\left(1 - \frac{e_j}{r_j}\right)$$

where k is the largest j such that $t_j < t$, $r_1 = N$ and $h_j = e_j/r_j$ is the estimate of the hazard, or conditional probability, of developing AD at time t_j given survival beyond time t_{j-1}. A perhaps more intuitive relationship between the hazard and survival functions involves $\hat{f}(t_j)$, the unconditional probability of failure at time t_j:

$$\hat{f}(t_j) = h_j \hat{S}(t_{j-1})$$

Thus the unconditional probability of failing at time t_j is the product of the conditional probability of failing at time t_j given survival beyond t_{j-1}, and the probability of surviving

beyond time t_{j-1}. The cumulative incidence function is

$$\hat{F}(t) = \sum_{j=1}^{k} \hat{f}(t_j)$$

where k is the largest j such that $t_j < t$, and the survival function is $\hat{S}(t) = 1 - \hat{F}(t)$. This recursive method for calculating the survival function and the cumulative incidence function begins at time t_1 and we define $\hat{S}(t_0) = \hat{S}(0) = 1$.

The traditional Kaplan–Meier method assumes that the time scale for failure times is time on study or some other function of calendar time. It can, however, be modified for use with a survival age time scale:

$$\hat{f}_A = h_A \hat{S}_{A-1} \tag{4}$$

where $\hat{S}_A$is the probability of surviving beyond age A. Notice that the hazard of developing the event at age A, h_A, is zero for ages less than $A_{\min}$, the youngest age at which the event occurs. In our case, the earliest diagnosis of AD was for a subject who was 70 and so $A_{\min} = 70$. Then the cumulative incidence at age A is

$$\hat{F}_A = \sum_{j=A_{\min}}^{A} \hat{f}_j = \sum_{j=A_{\min}}^{A} h_j \hat{S}_{j-1} \tag{5}$$

and the survival function is $\hat{S}_A = 1 - \hat{F}_A$. As was noted in Section 4.1, the size of the risk set is not necessarily non-decreasing; however, this does not impact the calculation of $\hat{S}_A$.

The *remaining lifetime risk* of failure from the event of interest is simply the cumulative incidence, $\hat{F}_{A_{\max}}$, where $A_{\max}$ is the maximum age. This method can easily be modified to condition on survival to a particular age, A_S, by setting $\hat{S}_A = 1$ for $A < A_S$, and by summing from $j = A_S$ instead of from $j = A_{\min}$. In this way, we can calculate the remaining lifetime risk of failure conditional on survival to age A_S. For example, we can calculate the remaining lifetime risk of developing AD for a cognitively intact 70-year-old or 75-year-old.

The variance of the estimated cumulative incidence at age A is given by Greenwood's formula [2, 28]:

$$\text{var}(\hat{F}_A) = \hat{S}_A^2 * \sum_{j=A_{\min}}^{A} \left(\frac{e_j}{r_j * (r_j - e_j)} \right)$$

95 per cent confidence limits can be constructed in the usual way as

$$\hat{F}_A \pm 1.96 \sqrt{\{\text{var}(\hat{F}_A)\}} \tag{6}$$

Note that in the calculation of $\hat{F}_A$ and its variance, deaths are treated as censored events. The estimate is not adjusted for the competing risk of death. For this reason, we term $\hat{F}_A$ the unadjusted cumulative incidence or UCI.

4.5. Cumulative incidence adjusted for competing risk (ACI)

We now return to the issue of competing risks. As discussed in Section 2.4, the unadjusted cumulative incidence, or UCI, overestimates the risk of actually developing AD. The method we

use to adjust the cumulative incidence of AD adjusted for the competing risk of death is described in detail by Gaynor *et al.* The adjustment is to the estimate of unconditional probability of failure, $\hat{f}_A$. In equation (4), the unconditional probability of failure at age A is the product of the hazard of failure at age A given survival to age (A-1). The estimate of the probability of survival to age (A-1) is based on an analysis in which deaths are censored and thus is an estimate of the probability of survival AD-free but not necessarily alive. A more appropriate condition is survival free of AD *and alive.* This probability may be obtained by performing a survival analysis (as described in Section 4.4) in which deaths are not censored but are counted as events of interest along with AD events. Using the notation in Section 4.1, the hazard of failing from either AD or death at age a is (c_a/r_a). The estimated survival probability from such an analysis is:

$$\hat{U}_A = 1 - \sum_{j=A_{\min}}^{A} (c_j/r_j)\,\hat{U}_{j-1}$$

Then we modify equation (4) as follows:

$$\hat{f}_A^* = h_A \hat{U}_{A-1} \tag{7}$$

The ACI, or cumulative incidence adjusted for the competing risk of death, is

$$\hat{F}_A^* = \sum_{j=A_{\min}}^{A} h_A \hat{U}_{A-1}. \tag{8}$$

The variance of the ACI can be estimated using a Taylor series linear expansion [12]:

$$\mathrm{SE}(\hat{F}_A^*) = \sqrt{\left\{ \sum_{j=A_{\min}}^{A} h_j \hat{U}_A \times \frac{(r_j - e_j)}{(e_j \times r_j)} + \sum_{1=A_{\min}}^{i-1} \frac{c_1}{r_1(r_1 - c_1)} + 2 \sum_{j=A_{\min}}^{A-1} \sum_{k=j+1}^{A} h_j \hat{U}_A \times h_k \hat{U}_A \left[\frac{-1}{r_j} + \sum_{1=A_{\min}}^{j-1} \frac{c_1}{r_1(r_1 - c_1)} \right] \right\}}$$

5. MACRO DESCRIPTION

In this section we describe our SAS macro *practical incidence estimators* (PIE) which provides estimates of age-specific incidence rates, crude and age-adjusted incidence rates, estimates of the unadjusted cumulative incidence and cumulative incidence rates adjusted for competing risk of death. The macro also provides the remaining lifetime risk of developing the event of interest conditional on survival to selected ages. We describe the macro parameters and the modules (sub-macros) it calls.

5.1. Preparing the data

The syntax used to call the SAS macro *practical incidence estimators* is as follows:

%macro PIE (IDS, minage, maxage, agegrpw, group, level1, level2, agefree, o1,o2);

The ten parameters are:

1. The input data set (IDS). This data set must contain the following variables and one observation per subject
   (i) Study identification number (*id*): unique identification numbers to distinguish one subject in the sample from another.
   (ii) Entry age (*entryage*): age, in years, at the beginning of the observation period.
   (iii) Survival age (*survage*): age, in years, in the last year of follow-up (for example, the year of failure from the event of interest or the year of censoring).
   (iv) Event status (*status*): an indicator variable coded 1 for subjects who develop the event of interest during the observation period (for example, develop AD during 1975–1998) and 0 for subjects who do not. (Subjects who die during the observation period are coded 0 with respect to event status.)
   (v) Any event status (*astatus*): an indicator variable coded 1 for subjects who fail due to the event of interest *or* the competing risk during the observation period (for example, develop AD or die during 1975–1998) and 0 for subjects who do not.
   (vi) Grouping variable (*group*): a variable that defines comparison groups of interest (for example, subject gender).
2. *minage*: minimum age at event onset.
3. *maxage*: maximum age at event onset.
4. *agegrpw*: the width of (or number of years in) each age interval used in the incidence tables.
5. *group*: the variable which defines comparison groups of interest with levels as in points 6 and 7.
6. *level* 1.
7. *level* 2.
8. *agefree*: age to which subjects are assumed to be free of the event (used in calculating future risk conditional on survival to age *agefree*).
9. *o*1: the name of the SAS data set with estimates of UCI and ACI derived from data in *level* 1 of the variable *group*.
10. *o*2: the name of the SAS data set with estimates of UCI and ACI derived from data in *level* 2 of the variable *group*.

The Framingham dementia cohort data set ('addata') is the IDS used in the PIE macro. Recall that each subject contributes one observation with the five requisite variables given above. The grouping variable we will consider here is *male* which is coded 1 for men and 0 for women. The constant parameter values are as follows:

minage = 70, the youngest age at diagnosis in our data;
maxage = 99, the oldest age at diagnosis in our data;
agegrpw = 5 requests 5-year age groups in the tables of one-year incidence;
group = male;
level 1 = 1;
level 2 = 0;
agefree = 70;
*o*1 = out1;
*o*2 = out0.

Thus our macro call is

%macro PIE (addata, 70, 99, 5, male, 1, 0, 70, out1, out0);

The PIE macro contains the following modules, each of which is described below:

%macro PIE (IDS, minage, maxage, agegrpw, group, level1, level2, agefree, o1, o2);

```
  %data1 (&IDS, PDS)
  %sdsmac (PDS, (&level1, &level2), SDS)
  title "&group = &level1";
  %sdsmac (PDS, (&level1), SDS1)
  title "&group = &level2";
  %sdsmac (PDS, (&level2), SDS2)
  title;
  %incid (SDS, I, 1)
  %incid (SDS1, i1, 0)
  %incid (SDS2, i2, 0)
  %aa2g (i1, i2)
  title "&group = &level1";
  %lr (SDS1, &e o1)
  title "&group = &level2";
  %lr (SDS2, &o2)
%mend;
```

5.2. Module one: creating a pooled data set

In the first module, a data set is created with one line per subject for each year at risk during the observation period. Subjects are considered at risk at a given age during the observation period as long as they are free of the event of interest at least until that age. Only ages that are between the specified minimum and maximum ages are included. For example, suppose a person enters the study at age 50 in 1975 and is still alive and free of AD in 1998. If we specify the minimum and maximum ages to be 70 and 99, respectively, this person contributes 4 years of risk, starting in 1995 at age 70, to 1998 at age 73. The pooled data set excludes those years when subjects were free of the event of interest but were observed at ages less than the specified minimum age at event onset. With respect to our example subject, the first 20 years of observation, when the subject was aged 50–69, are excluded from the pooled data set. The pooled data set also excludes ages at risk that occurred before the observation period.

The calculation of one-year incidence rates uses a weighting scheme similar to the strategy employed in the construction of actuarial estimates (Section 4.1). In most cases, the last year of follow-up for subjects who are censored free of AD is assigned a weight of 0.5. However, subjects who are censored at the maximum age to be considered but who are known to have survived free of AD to an older age are assigned weights of 1.0 in their last year of follow-up. All other observations are assigned weights of 1.0.

The module *data* 1 is called as follows, where *in* and *out* are the input and pooled data sets, respectively. The PIE macro invokes this module to produce the pooled data set *PDS* based on the initial data set *IDS*.

```
%macro data1 (in, out);
  data &out;set &in;
    keep id age status astatus weight &group;
    if survage gt &maxage then do;
      survage = &maxage; status = 0; astatus = 0; full = 1;
    end;
    start = max(entryage, &minage);
    stop = survage-1;
    age = survage;
    if status eq 0 and full ne 1 then weight = 0.5;else weight = 1;
    output;
    do age = start to stop;
      status = 0; astatus = 0; weight = 1; output;
    end;
    run;
%mend;
```

5.3. Module two: creating a summary data set for each age

In the second module, the pooled data set (created in module one) is summarized for each year of age. Specifically, for each year of age, A, between the minimum and maximum ages, the module outputs r_A, the number of subjects in the risk set, e_A, the number of events of interest, c_A, the total number of events (including the competing risk) and w_A, the weighted number of person years at risk (see Section 4.1). In addition, the module outputs the sum over all ages of each of r, e, c and w.

The PIE macro invokes the module three times, each time inputting the pooled data set *PDS* created in module one. The first call produces a summary data set, *SDS*, for all subjects combined. The next two calls of the module produce summary data sets *SDS*1 and *SDS*2, for the two levels of the comparison group of interest. In our example, SDS is a summary data set for men and women combined, and *SDS*1 and *SDS*2 are summary data sets for men and women, respectively.

```
%macro sdsmac (in, level, out);
  proc means noprint n sum data = &in;
    where (&group in &level);
    class age;
    var status astatus weight;
    output out = t n = r sum = e c w;
  run;
  data &out; set t;
    if age eq . then age = 999;
    proc sort; by age; run;
%mend;
```

5.4. Module three: computing age-specific incidence rates

In the third module, each summary data set output in module two (*SDS*, *SDS*1 and *SDS*2) is used to calculate one-year incidence rates. First, ages are collapsed into intervals of specified width (for example, if *minage* = 65, *maxage* = 94, and *agegrpw* = 5, ages are collapsed into 5-year

intervals: 65–69, 70–74, 75–79, ..., 90–94). The number of events of interest (e_A) and the number of weighted person-years (w_A) are summed over ages within each age group G and over all ages. Equations (1) and (2) are applied to obtain age group-specific one-year incidence rates per 1000 person years (see Section 4.2).

The age-specific and crude overall rates are output in the data set *out*. The PIE macro invokes this module three times, once for all subjects combined, and once each for the two levels of the grouping variable. The output data sets corresponding to the summary data sets *SDS*, *SDS*1 and *SDS*2 are called *I*, *I*1 and *I*2, respectively. In addition to the age group-specific numbers of events, weighted person-years, and rates, each observation in the data set *out* contains the total number of weighted person-years; these will all be used in the next module to calculate age-adjusted rates. The next module inputs the two data sets, *I*1 and *I*2, corresponding to the two levels of the grouping variable. Note that age group is assigned a format that requires definition before the macro is invoked.

The incidence rates may be printed in this module by setting the parameter p equal to 1 when calling the macro. Level specific rates are automatically printed out in the next module and need not be printed here; however, overall incidence rates should be printed here.

```
%macro incid (in, out, p);
  data temp; set &in;
    if &minage le age le &maxage;
    agegroup = floor((age-&minage)/&agegrpw);
  run;
  proc means noprint sum; class agegroup;
    var e w;
    output out = t sum = events wpy;
  run;
  data tot tt;set t;
    wtr = 1000*events/wpy;
    if agegroup eq . then do; agegroup = 88; output tot; end;
    output tt;
  run;
  proc sort; by agegroup; run;
  data &out;
    format agegroup agf.;
    if _n_ eq 1 then set tot (rename = (wpy = twpy));
    set tt;
    last = (floor((&maxage- &minage)/ &agegrpw));
  run;
  %if &p eq 1 %then %do;
    proc print; var agegroup events wpy wtr; run;
  %end;
%mend;
```

5.5. Module four: computing age-adjusted incidence rates

In the fourth module, age-adjusted rates are produced using direct standardization of the level-specific incidence rates computed in module three. The standardization is to the distribution

of person-years in the combined group. The module calculates the overall proportion of weighted person-years in each age group and applies this to the level-specific rates to obtain age-adjusted rates (see Section 4.3). The data sets *I*1 and *I*2 output in module three are input here; this module produces printed output only.

```
%macro aa2g (in1, in2);
  data a;
    merge &in1 (rename = (events = e&level1 wpy = wpy&level1
                          twpy = twpy&level1 wtr = rate&level 1))
          &in2 (rename = (events = e&level2 wpy = wpy&level2
                          twpy = twpy&level2 wtr = rate&level2));
    by agegroup;
  data b;set a (drop = rate&level1 rate&level2);
    if agegroup ne 88;
    keep agegroup rate&level1 rate&level2;
    atwpy = sum (of wpy&level1, wpy&level2);
    gtwpy = sum (of twpy&level1, twpy&level2);
    wtwpy = atwpy/gtwpy;
    retain rate&level1 rate&level2;
    rate&level1 =
      sum (of rate&level1, 1000* (e&level1/wpy&level1)*wtwpy);
    rate&level2 =
      sum (of rate&level2, 1000* (e&level2/wpy&level2)*wtwpy);
    if age group eq last then do;
      agegroup = 99; output;
      end;
    data c;
      set a b;
    run;
    proc print noobs;
    var agegroup e&level1 wpy&level1 rate&level1
                  e&level2 wpy&level2 rate&level2; run;
%mend;
```

5.6. Module five: computing unadjusted cumulative incidence (UCI) *and cumulative incidence adjusted for competing risk* (ACI)

This module produces point and interval estimates of unadjusted cumulative incidence (UCI) and cumulative incidence adjusted for the competing risk of death (ACI), indexed by age (in years). All estimates are conditional on survival (alive and event-free) until age *agefree* specified in the PIE macro call. The PIE macro invokes this module once for each level of the grouping variable, inputting in each case the summary data set produced by module two (for example, *SDS*1 for level 1).

This module produces a printed table of estimated UCI along with 95 per cent confidence limits, and a printed table of estimated ACI along with 95 per cent confidence limits, indexed in each case by the number of years at risk. The numbers of years at risk selected for

printing are a function of the constant *agegrpw* specified in the PIE macro call. For example, if *agegrpw* = 5, PIE produces 5-year risk, 10-year risk, 15-year risk, and so on. The entire set of estimates and their standard errors, for each year of age (or number of years since *agefree*), are output to the data set *sendout*. These data are intended for export to graphical software to produce plots.

The module estimates, at age A, the hazard for the event of interest, $h_A = (e_A/w_A)$, and the hazard for all events (the event of interest or the competing event), (c_A/w_A). Conditional on survival to *agefree*, the UCI (*cf&agefree*) and ACI (*cfstar&agefree*) are estimated in this module according to equations (4) and (5) in Section 4.4, and equations (7) and (8) in Section 4.5.

```
%macro lr (in, sendout);
  data out; set &in;
    if (&agefree le age le &maxage);
    array a (*) a&agefree-a&maxage; array b (*) b&agefree-b&maxage;
    array fstara (*) fstar&agefree-fstar&maxage;
    index = age-&agefree + 1;
    retain cs cvlp cfa csa cov v cf&agefree cfstar&agefree
           a&agefree-a&maxage b&agefree-b&maxage
           fstar&agefree-fstar&maxage;
    if age eq &agefree then do;
      cs = 1; cvlp = 0; cfa = 0; csa = 1; cov = 0; v = 0;
      cf&agefree = 0; cfstar&agefree = 0;
    end;

  ** survival estimates***;
    h = e/r; ha = c/r;
    f = h*cs; cf&agefree = cf&agefree + f; cs = 1-cf&agefree;
    fa = ha*csa; fstar = h*csa;
    cfa = cfa + fa; csa = 1-cfa;
    cfstar&agefree = cfstar&agefree + fstar;
  **standard errors***;
    cvlp = cvlp + e/(r*(r-e));
    if age ge &agefree then do;
      a(index) = (-1/r); fstara(index) = fstar;
    end;
    if age eq &agefree then b&agefree = c/(r*(r-c));
    if age gt &agefree then
      b(index) = b(index-1) + (c/(r*(r-c)));
    cc = 0;
    if age ge (&agefree + 2) then
    do k = 2 to (index-1);
      cc = cc + (fstara(k)*(a(k) + b(k-1)));
    end;
    cov = cov + fstar*cc;
    if (age ge (&agefree + 1)) then bb = b(index-1);
    else bb = 0;
```

```
    if (e gt 0) then
      v = v + (fstar**2)*(((r-e)/(e*r)) + bb);
    secf&agefree = sqrt((cs**2)*cvlp);
    secff&agefree = sqrt(v + 2*cov);
    keep age r e c cf&agefree secf&agefree
         cfstar&agefree secff&agefree;
  run;

  data s1; set out;
    keep years cf&agefree secf&agefree cfstar&agefree secff&agefree
         lcl ucl lclstar uclstar;
    cf&agefree = cf&agefree*100;
    secf&agefree = secf&agefree*100;
    cfstar&agefree = cfstar&agefree*100;
    secff&agefree = secff&agefree*100;
    years = (age- &agefree + 1);
    lcl = max(0, cf&agefree-1.96*secf&agefree);
    ucl = min(100, cf&agefree + 1.96*secf&agefree);
    lclstar = max(0, cfstar&agefree-1.96*secf&agefree);
    uclstar = min(100,cfstar&agefree + 1.96*secff&agefree);
  run;

  data &sendout; set s1;
     keep years cf&agefree secf&agefree
                cfstar&agefree secff&agefree;
    run;

    data s2; set s1;
      if ((years/&agegrpw eq floor (years/&agegrpw)));
        format years yf.;
  run;
  title2 'Unadjusted Cumulative Incidence';
  title3 'With 95% Confidence Limits';
  proc print; id years; var cf&agefree lcl ucl; run;
  title2 'Cumulative Incidence, Adjusted for Competing Risk of Death';
  title3 'With 95% Confidence Limits';
  proc print; id years; var cfstar&agefree lclstar uclstar; run;
  title; title2; title3;
%mend;
```

6. RESULTS

Over the course of the follow-up period (from 1975 to 1998), 2441 subjects (972 men and 1469 women) contributed at least one year between age 70 and 99, with a total of 27044 person-years. During this period, 172 subjects developed AD, 1336 died without developing AD and 933 were censored. The earliest age at AD *onset* was 65 years; however, the earliest age at AD *diagnosis* was 70 years.

We applied the PIE macro to these data twice, conditioning first on survival free of AD to age 70 and then on survival free of AD to age 75. The third and fourth modules in which age-specific and age-adjusted one-year incidence rates are calculated were only invoked in the first application.

6.1. Age-specific and age-adjusted one-year incidence

The first call to module three produces the following table which includes age group-specific one-year incidence rates of AD per 1000 person-years (*wpr*), along with the numbers of AD cases and person-years (*events* and *wpy*, respectively). The table provides rates for 5-year age groups (*agegrpw* = 5) from age 70 (*minage* = 70) to age 99 (*maxage* = 99), and the crude rate over ages 70–99. The one-year rate per 1000 person-years increases from 0.8 in subjects aged 70–74 to 51.4 per 1000 person-years in subjects aged 94-99. The greatest increase is from 7.1 in subjects aged 80–84 to 21.0 in subjects aged 85–89. Of particular interest is that the rate continues to increase even in subjects over 90 years of age. Among subjects aged 70–99, there were 172 cases of AD in the 27044 person-years, and the one-year rate of incident AD was 6.4 per 1000 person years.

OBS	AGEGROUP	EVENTS	WPY	WTR
1	70-74	7	8967.0	0.7806
2	75-79	24	8585.5	2.7954
3	80-84	40	5652.0	7.0771
4	85-89	58	2756.0	21.0450
5	90-94	34	908.5	37.4243
6	95-99	9	175.0	51.4286
7	CR 70-99	172	27044.0	6.3600

The following gender-specific incidence rates are produced by the fourth module and are summarized in Table I. Age group-specific numbers of events (*e*), numbers of person-years (*wpy*), and one-year incidence rates (*rate*), are given for each level of the grouping variable, *male* (coded 1 for men and 0 for women). The one-year incidence of AD increases with age in both men and women (except for a small decrease in men from 85–89 to 90–94). The crude one-year incidence rates per 1000 person-years are 4.4 and 7.5 for men and women, respectively. Direct standardization to the overall age distribution yields age-adjusted one-year incidence rates per 1000 person-years of 5.0 and 6.9 for men and women, respectively.

AGEGROUP	E1	WPY1	RATE1	E0	WPY0	RATE0
70-74	5	3597.5	1.3899	2	5369.5	0.3725
75-79	8	3232.0	2.4752	16	5353.5	2.9887
80-84	13	1984.5	6.5508	27	3667.5	7.3620
85-89	12	804.5	14.9161	46	1951.5	23.5716
90-94	3	231.0	12.9870	31	677.5	45.7565
95-99	2	30.5	65.5738	7	144.5	48.4429
CR 70-99	43	9880.0	4.3522	129	17164.0	7.5157
AA 70-99	.	.	4.9964	.	.	6.8636

The numbers of events in some of the gender-specific age groups are small enough to warrant combining age groups and we invoked the PIE macro with *agregrpw* set to 10 to produce the following table of gender-specific incidence rates in 10-year age groups (summarized in Table II).

Table I. Gender specific one-year incidence of AD per 1000 person-years: age specific, crude and age-adjusted rates (5-year age groups).

Age group	Men		Women	
	Number of events	Incidence rate	Number of events	Incidence rate
70–74	5	1.4	2	0.4
75–79	8	2.5	16	3.0
80–84	13	6.5	27	7.4
85–89	12	14.8	46	23.6
90–94	3	13.0	31	45.8
95–99	2	65.6	7	48.4
Crude rate 70–99	43	4.4	129	7.5
Age-adjusted rate 70–99		5.0		6.9

Table II. Gender specific one-year incidence of AD per 1000 person-years. age specific, crude and age-adjusted rates (10-year age groups).

Age group	Men		Women	
	Number of events	Incidence rate	Number of events	Incidence rate
70–79	13	1.9	18	1.7
80–89	25	9.0	73	13.0
90–99	5	19.1	38	46.2
Crude rate 70–99	43	4.4	129	7.5
Age-adjusted rate 70–99		4.8		7.0

The one-year incidence rates in men and women aged 70–79 are similar, and both increase with age. However, the rates increase over the three decades much more sharply in women (from 1.7 to 13.0 to 46.2) than in men (from 1.9 to 9.0 to 19.1).

AGEGROUP	E1	WPY1	RATE1	E0	WPY0	RATE0
70-79	13	6829.5	1.9035	18	10723	1.6786
80-89	25	2789.0	8.9638	73	5619	12.9916
90-99	5	261.5	19.1205	38	822	46.2287
CR 70-99	43	9880.0	4.3522	129	17164	7.5157
AA 70-99	.	.	4.7883	.	.	6.9807

6.2. Unadjusted cumulative incidence and cumulative incidence adjusted for the competing risk of death in subjects who survived to age 70 free of AD

The following gender-specific unadjusted and adjusted cumulative incidence rates for 5, 10, 15, 20, 25 and 30 year risks of developing AD are output in module five (shown following) along with 95 per cent confidence intervals. Table III summarizes the output. All rates are conditional on survival free of AD to age 70 and risks are calculated from age 70. We define the

remaining lifetime risk of developing AD to be the risk of developing AD by age 99 or the 30 year risk.

The risk of developing AD based on unadjusted cumulative incidence is similar for men and women, with the same sharp increase from 15 year to 20 year risk that we observed in the one-year incidence rates from the 80–84 to the 85–89 year age group. The remaining lifetime risks of developing AD for subjects who have survived free of AD to age 70 are 43.0 per cent (95 per cent CI: 14.4–71.6 per cent) for men and 48.3 per cent (95 per cent CI: 35.6–61.1 per cent) for women.

male = 1 Event-free to age 70
Unadjusted Cumulative Incidence
With 95% Confidence Limits

YEARS	CF70	LCL	UCL
5 Yr Risk	0.6787	0.0858	1.2715
10 Yr Risk	1.8641	0.8576	2.8705
15 Yr Risk	5.2220	3.1546	7.2895
20 Yr Risk	11.9038	7.7071	16.1005
25 Yr Risk	20.8547	10.3412	31.3683
30 Yr Risk	43.0154	14.3861	71.6448

male = 1 Event-free to age 70
Cumulative Incidence, Adjusted for Competing Risk of Death
With 95% Confidence Limits

YEARS	CFSTAR70	LCLSTAR	UCLSTAR
5 Yr Risk	0.64603	0.05317	1.21078
10 Yr Risk	1.57913	0.57269	2.43194
15 Yr Risk	3.36349	1.29603	4.63990
20 Yr Risk	5.50220	1.30555	7.22862
25 Yr Risk	6.34228	0.00000	8.28972
30 Yr Risk	7.05981	0.00000	9.20926

male = 0 Event-free to age 70
Unadjusted Cumulative Incidence
With 95% Confidence Limits

YEARS	CF70	LCL	UCL
5 Yr Risk	0.1817	0.0000	0.4333
10 Yr Risk	1.6626	0.8999	2.4253
15 Yr Risk	5.1636	3.6631	6.6641
20 Yr Risk	15.3778	12.2078	18.5478
25 Yr Risk	32.5752	26.3706	38.7798
30 Yr Risk	48.3038	35.5570	61.0506

male = 0 Event-free to age 70
Cumulative Incidence, Adjusted for Competing Risk of Death
With 95% Confidence Limits

YEARS	CFSTAR70	LCLSTAR	UCLSTAR
5 Yr Risk	0.1767	0.00000	0.4214
10 Yr Risk	1.4685	0.70581	2.1424
15 Yr Risk	4.0255	2.52501	5.1833
20 Yr Risk	9.5079	6.33786	11.3954
25 Yr Risk	14.7764	8.57176	17.2786
30 Yr Risk	16.5520	3.80514	19.2941

Table III. Estimates of gender-specific risk of AD conditional on survival free of AD to age 70.

Years	Men		Women	
	UCI* (95% CI)	ACI† (95% CI)	UCI* (95% CI)	ACI† (95% CI)
5 year risk	0.7 (0.1, 1.3)	0.6 (0.1, 1.2)	0.2 (0.0, 0.4)	0.2 (0.0, 0.4)
10 year risk	1.8 (0.9, 2.9)	1.6 (0.6, 2.4)	1.7 (0.9, 2.4)	1.5 (0.7, 2.1)
15 year risk	5.2 (3.2, 7.3)	3.4 (1.3, 4.6)	5.2 (3.7, 6.7)	4.0 (2.5, 5.2)
20 year risk	11.9 (7.7, 16.1)	5.5 (1.3, 7.2)	15.4 (12.2, 18.5)	9.5 (6.3, 11.4)
25 year risk	20.9 (10.3, 31.4)	6.3 (0.0, 8.3)	32.6 (26.4, 38.8)	14.8 (8.6, 17.3)
Lifetime risk	43.0 (14.4, 71.6)	7.1 (0.0, 9.2)	48.3 (35.6, 61.1)	16.6 (3.8, 19.3)

* Unadjusted cumulative incidence.
† Cumulative incidence, adjusted for competing risk of death.

Table IV. Estimates of gender-specific risk of AD, conditional on survival free of AD to age 75.

Years	Men		Women	
	UCI* (95% CI)	ACI† (95% CI)	UCI* (95% CI)	ACI† (95% CI)
5 year risk	1.2 (0.4, 2.0)	1.1 (0.3, 1.8)	1.5 (0.8, 2.2)	1.4 (0.7, 2.1)
10 year risk	4.6 (2.6, 6.6)	3.2 (1.2, 4.5)	5.0 (3.5, 6.5)	4.1 (2.6, 5.3)
15 year risk	11.3 (7.1, 15.5)	5.6 (1.4, 7.5)	15.2 (12.1, 18.4)	10.0 (6.8, 12.0)
20 year risk	20.3 (9.7, 30.9)	6.6 (0.0, 8.8)	32.5 (26.2, 38.7)	15.6 (9.4, 18.3)
Lifetime risk	42.6 (13.8, 71.4)	7.4 (0.0, 9.9)	48.2 (35.4, 61.0)	17.5 (4.7, 20.4)

* Unadjusted cumulative incidence.
† Cumulative incidence, adjusted for competing risk of death.

Adjusting for the competing risk of death substantially reduces the estimates of cumulative incidence of developing AD in both men and women. The adjustment has a more pronounced impact on 20, 25 and 30 year risks because mortality is much higher as age increases to 85 and older. For example, the UCI estimate of remaining lifetime risk is 43.0 per cent for men compared to 7.1 per cent after adjustment for the competing risk of death. Among women, the UCI estimate of remaining lifetime risk is 48.3 per cent as compared to 16.6 per cent after adjustment for the competing risk of death.

Module five produces two data sets (*out*1 and *out*0) with the unadjusted (UCI) and adjusted (ACI) cumulative incidence estimates for men and women, respectively. The four estimated cumulative incidence curves are plotted in Figure 3. Figure 3(*a*) presents the *UCI* curves for men and women. Figure 3(*b*) compares the UCI and ACI curves for women. It is clear that the UCI and ACI curves are similar for about 15 years when the UCI curves increase sharply and continue to increase to age 99 (30 year risk). The ACI curves continue to increase but at a much slower rate. Figure 3(*c*) compares the ACI curves for men and women. Cumulative incidence of AD is similar in men and women for about 15 years and then is higher in women than in men. In particular, after 20 years (at age 90), the ACI in women is about twice as high as that in men.

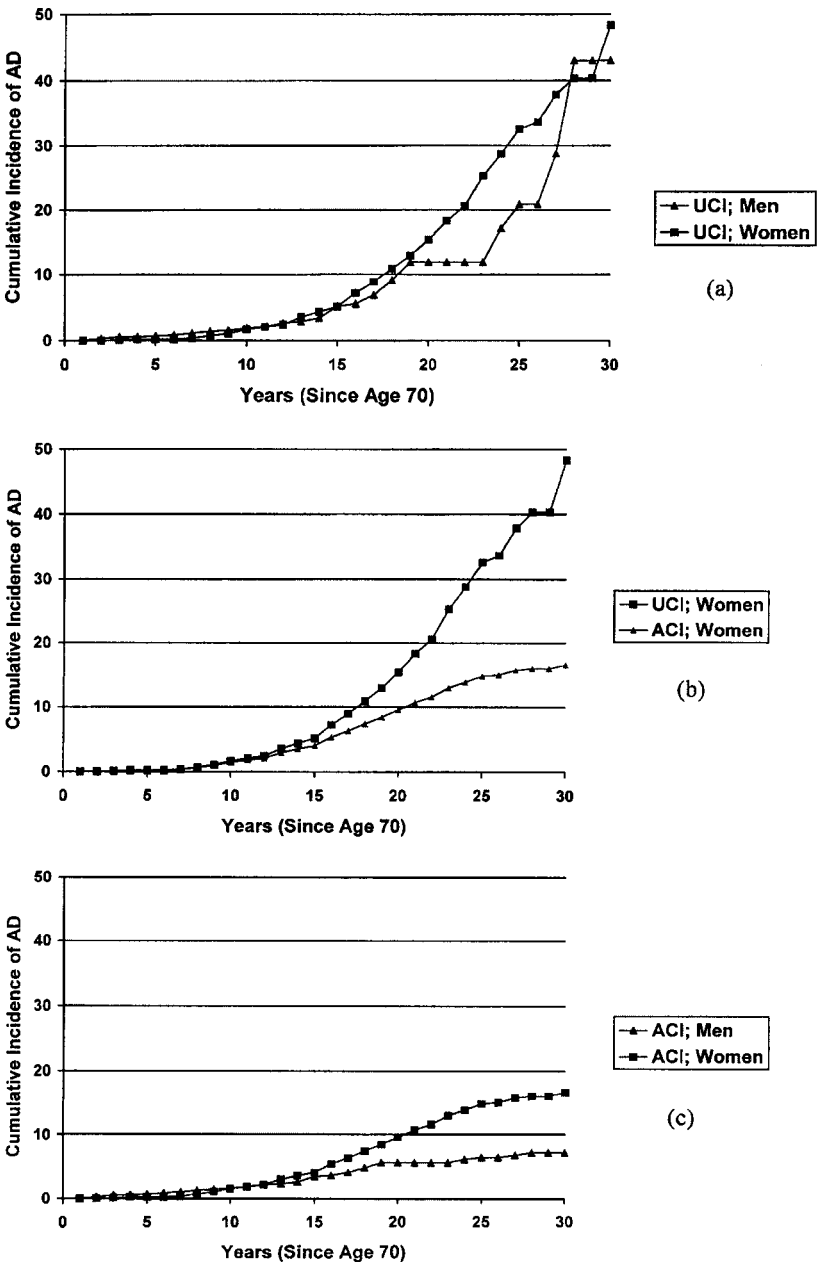


Figure 3. (*a*) Estimates of gender-specific unadjusted cumulative incidence of Alzheimer's disease conditional on survival dementia-free to age 70 years. (*b*) Estimates of unadjusted cumulative incidence of Alzheimer's disease and cumulative incidence of Alzheimer's disease adjusted for the competing risk of death dementia-free to age 70 years in women who survived. (*c*) Estimates of gender-specific cumulative incidence of Alzheimer's disease adjusted for the competing risk of death conditional on survival dementia-free to age 70 years.

7. DISCUSSION

The application of traditional statistical techniques of risk-factor epidemiology in the very old can lead to erroneous conclusions due to special features of the population. The most problematic issue is the steep age-related increase in both the mortality rate and the incidence of degenerative diseases such as Alzheimer's disease. In 1992 and 1993, we published data on the prevalence and incidence of all-cause dementia and Alzheimer's disease (AD) in the original Framingham cohort [29, 30]. As this population aged, we addressed the impact of mortality on cumulative risk estimates within this cohort and showed that earlier cumulative incidence projections as high as 49.6 per cent [31] did not reflect the true disease experience of an elderly cohort.

Techniques for estimating incidence of disease are described in detail in texts on epidemiologic methods [8] and biostatistics [10, 11], particularly texts that cover survival analysis methods [2]. Statistical software packages such as SAS are well equipped to produce many of the traditional measures. However, estimates of incidence derived from data collected in certain scenarios (for example, in a very old population) are not easy to obtain using currently available software.

Suppose that data are structured with one observation per subject containing disease status and survival time. SAS *Proc Freq* summarizes data into frequency histograms or $R \times C$ tables and will provide tables of numbers (and per cents) of incident events among all subjects or stratified by a grouping variable. The incidence measures calculated in these tables assume that all subjects have been observed over the same period of time and do not account for right censoring. This is clearly not the case when subjects are assessed for incident disease over a long period of time. The useful descriptive summary of incidence in this scenario relies on the person-years approach in which incidence is reported as a rate per specified number of person-years. It is relatively easy to expand the data to a pooled data set with one observation per person-year or person-exam. *Proc Freq* applied to the pooled data will produce the appropriate incidence measures, and will even produce tables of age-specific incidence rates with some additional programming. The calculation of age-adjusted rates using direct standardization requires yet more custom programming. Incidence rates per person-years, age-specific incidence rates, and age-adjusted incidence rates are useful descriptive statistics and should be readily available.

Another useful descriptive summary of survival data is cumulative incidence over a specified time scale. *Proc Lifetest* provides life table estimates of cumulative incidence and allows the selection of the actuarial method or the Kaplan–Meier product-limit method. In addition, several tests are available for comparing survival curves among levels of a grouping variable. *Proc Phreg* performs proportional hazards regression using Cox's partial likelihood method; estimates of cumulative survival (and incidence) can be output for specified values of covariates (that is, for levels of a grouping variable). *Proc Phreg* supports the inclusion of time-varying covariates, and performs conditional logistic regression, among other options. Its more recent version also allows the user to specify a *left-truncation* variable to ensure that only time that is observed during the study period is included. Thus, *Proc Phreg* can be used to produce the unadjusted cumulative incidence (UCI) where the survival time variable is survival age and the left-truncation variable is age at entry.

The concept of remaining lifetime risk relies on the concept of remaining lifetime and the calculation of remaining lifetime risk should be derived using cumulative incidence adjusted for the competing risk of death. Estimates of the ACI are not produced by any of the survival analysis procedures in software packages such as SAS.

The PIE macro produces crude and age-specific incidence rates, overall and stratified by the levels of a grouping variable. In addition, it produces age-adjusted rates using direct standardization to the combined group. The user determines the width of the age groups. We have presented the macro for use with grouping variables with two levels. We do, however, have other modules to be invoked when there are three or more levels, and the PIE macro can be modified to allow the user to specify the number of levels of the grouping variable.

The PIE macro produces both the UCI and the ACI, and their respective standard errors, both in table form and in an output data set for graphing. The macro is designed for use with survival age as the time variable, and with age at entry into the study as the left-truncation variable; however, calendar time can be substituted for the survival time variable and the left-truncation variable can simply be set to zero. Perhaps the PIE macro's most useful product is the estimate of remaining lifetime risk conditional on survival alive and event-free to a specified age. The macro has the flexibility to produce estimates of remaining lifetime risk conditional on survival event-free to various user-specified ages. This allows the user to investigate the impact of increasing age on the estimate of remaining lifetime risk of disease.

We present the PIE macro using the example of Alzheimer's disease incidence data collected in the Framingham Study. These data are unique in that we use discrete years instead of a continuous time scale; the PIE macro is designed for this time scale.

Our estimates of AD incidence may differ from those in the literature. First, we were able to establish a cohort of participants who were determined to be dementia-free in 1975 and have been observed prospectively since that time. Survival time is censored conservatively at the last time the participant was actually assessed and classified as cognitively intact. Second, we use specific, objective criteria to determine the year of AD diagnosis instead of estimating the year of AD onset which can be very subjective and can vary substantially depending on the source of historical information. The youngest AD case in our analyses has a diagnosis age of 70 years; the estimated age at onset would, of course, be younger.

The PIE macro presented in this paper provides an estimate of the cumulative incidence adjusted for the competing risk of death along with more standard measures of disease incidence. Our current research includes the development of methods to formally compare ACI curves among levels of a grouping variable (similar to the logrank test procedure in *Proc Lifetest*), and to investigate the effect of risk factors on the ACI (as *Proc Phreg* provides for the UCI).

REFERENCES

1. D'Agostino, RB, Kannel WB. Epidemiological background and design: The Framingham Study. Proceedings of the American Statistical Association Sesquicentennial Invited Paper Sessions, American Statistical Association, Alexandria, VA, 1989.
2. Collett D. *Modelling Survival Data in Medical Research.* Chapman and Hall: 1994.
3. Elandt-Johnson RC, Johnson NL. *Survival Models and Data Analysis.* Wiley: New York, 1980.
4. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data.* Wiley: New York, 1980.
5. Marubini M, Valsecchi MG. *Analysing Survival Data from Clinical Trials and Observational Studies.* Wiley: New York, 1995.
6. Cantor AB. *Extending SAS® Survival Analysis Techniques for Medical Research,* SAS Institute Inc: Cary, NC, 1997.
7. Allison PD. *Survival Analysis Using the SAS® System: A Practical Guide.* SAS Institute Inc: Cary, NC, 1995.
8. Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic Research: Principles and Quantitative Methods.* Van Nostrand Reinhold: New York, 1982.
9. Miettinen OS. *Theoretical Epidemiology.* Wiley: New York, 1995.
10. Rosner B. *Fundamentals of Biostatistics,* 5th edn. Duxbury Press: Pacific Grove, CA, 2000.

11. Fleiss JL. *Statistical Methods for Rates and Proportions*, 2nd edn. Wiley: New York, 1981.
12. Gaynor JJ, Feuer EJ, Tan C, Wu DH, Little CR, Straus DJ, Clarkson BD, Brennan MF. On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. *Journal of the American Statistical Association* 1993; **88**:400–409.
13. Gooley TA, Leisenring W, Crowley J, Storer BE. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in Medicine* 1999; **18**(6):695–706.
14. Benechou J, Gail MH. Estimates of absolute cause-specific risk in cohort studies. *Biometrics* 1990; **46**:813–826.
15. Pepe MS. Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association* 1991; **86**(415):770–778.
16. Pepe MS, Mori M. Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data. *Statistics in Medicine* 1993; **12**:737–751.
17. Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. *Biometrics* 1978; **34**:541–554.
18. Seshadri S, Wolf PA, Beiser A, Au R, McNulty K, White R, D'Agostino R. Lifetime risk of dementia and Alzheimer's disease: the impact of mortality on risk estimates in the Framingham Study. *Neurology* 1997; **49**:1498–1504.
19. Cox DR, Oakes D. *Analysis of Survival Data.* Chapman & Hall: 1984.
20. Korn EL, Graubard, BI, Midthune D. Time-to-event analysis of longitudinal follow-up of a survey: choice of time scale. *American Journal of Epidemiology* 1997; **145**(1):72–80.
21. Kaplan EL, Meier P. Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**:457–481.
22. Lloyd-Jones DM, Larson MG, Beiser A, D'Agostino RB, Kannel WB, Murabito JM, Vasan RS, Benjamin EJ, Levy D. Lifetime risk of developing congestive heart failure. *Circulation* 1999; **100**:1–396.
23. Framer ME, White LR, Kittner SJ, *et al.* Neuropsychological test performance in Framingham: a descriptive study. *Psychological Report* 1987; **60**:1023–1040.
24. Folstein MF, Folstein SE, McHugh PR. Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* 1975; **12**:189–198.
25. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*, 3rd edn, Rev.: DSM-IIIR. American Psychiatric Association: Washington, DC, 1987.
26. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn: DSM-IV. American Psychiatric Association: Washington DC, 1994.
27. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* 1984; **34**:939–944.
28. Greenwood M. The natural duration of cancer. Reports on Public Health and Medical Subjects, 33 Her Majesty's Stationary Office, London, 1926, 1–26.
29. Bachman DL, Wolf PA, Linn R. *et al.* Prevalence of dementia and probable senile dementia of the Alzheimer type in the Framingham Study. *Neurology* 1992; **42**:115–119.
30. Bachman DL, Wolf PA, Linn R, *et al.* Incidence of dementia and probable Alzheimer's disease in a general population: the Framingham Study. *Neurology* 1993; **43**:515–519.
31. Hebert LE, Scherr PA, Beckett LA, *et al.* Age-specific incidence of Alzheimer's Disease in a community population. *Journal of the American Medical Association*, 1995; **273**:1354–1359.

TUTORIAL IN BIOSTATISTICS

The applications of capture-recapture models to epidemiological data

Anne Chao[1,*,†], P. K. Tsay[1], Sheng-Hsiang Lin[1], Wen-Yi Shau[2] and Day-Yu Chao[3]

[1]*Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan*
[2]*Graduate Institute of Clinical Medicine, National Taiwan University, Taipei, Taiwan*
[3]*Graduate Institute of Epidemiology, National Taiwan University, Taipei, Taiwan*

SUMMARY

Capture-recapture methodology, originally developed for estimating demographic parameters of animal populations, has been applied to human populations. This tutorial reviews various closed capture-recapture models which are applicable to ascertainment data for estimating the size of a target population based on several incomplete lists of individuals. Most epidemiological approaches merging different lists and eliminating duplicate cases are likely to be biased downwards. That is, the final merged list misses those who are in the population but were not ascertained in any of the lists. If there are no matching errors, then the duplicate information collected from a capture-recapture experiment can be used to estimate the number of missed under proper assumptions. Three approaches and their associated estimation procedures are introduced: ecological models; log-linear models, and the sample coverage approach. Each approach has its unique way of incorporating two types of source dependencies: local (list) dependence and dependence due to heterogeneity. An interactive program, CARE (for capture-recapture) developed by the authors is demonstrated using four real data sets. One set of data deals with infection by the acute hepatitis A virus in an outbreak in Taiwan; the other three sets are ascertainment data on diabetes, spina bifida and infants' congenital anomaly discussed in the literature. These data sets provide examples to show the usefulness of the capture-recapture method in correcting for under-ascertainment. The limitations of the methodology and some cautionary remarks are also discussed.

1. INTRODUCTION

The purpose of many epidemiological surveillance studies is to estimate the size of a population by merging several incomplete lists of names in the target population. Some examples are as follows:

1. An outbreak of the hepatitis A virus (HAV) occurred in and around a college in northern Taiwan from April to July 1995 [1]. Cases of students in that college were

* Correspondence to: Anne Chao, Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043, Taiwan
† E-mail: chao@stat.nthu.edu.tw

Contract/grant sponsor: National Science Council of Taiwan; contract/grant numbers: NSC 87-2118-M007-101, NSC89-2118-M007-006

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
© 2004 John Wiley & Sons, Ltd. ISBN: 0-470-02365-1

ascertained by three sources: (i) P-list, records based on a serum test taken by the Institute of Preventive Medicine, Department of Health of Taiwan – there were 135 identified cases; (ii) Q-list, local hospital records reported by the National Quarantine Service – 122 cases were found; (iii) E-list, records collected by epidemiologists – there were 126 cases. Merging the three lists by eliminating duplicate records resulted in 271 ascertained cases. This data set has the advantage of a known true number of infected because a screen serological check for all students was conducted after the three surveys. In Section 5, we use this data set to show the need of correction for undercount.

2. Hook *et al.* [2] and Regal and Hook [3] presented a data set on spina bifida collected in New York State from 1969–1974. Three lists were collected from birth certificates, death certificates and medical rehabilitation files. There were 513, 207 and 188, cases, respectively, on the three lists and in total 626 ascertained cases. A method of estimating the number of missed cases was discussed by the above authors to assess the completeness of the survey and to accurately estimate the prevalence rate.
3. Bruno *et al.* [4] collected a data set on diabetes in a community in Italy based on the following four records: diabetic clinic and/or family physician visits (1754 cases); hospital discharges (452 cases); prescriptions (1135 cases), and purchases of regent strips and insulin syringes (173 cases). A total of 2069 cases were identified. Despite the active identification, Bruno *et al.* concluded that there were still some people who could not be identified. The purpose was then to estimate the number of missing diabetes patients and to adjust for undercount.
4. Wittes *et al.* [5] and Fienberg [6] analysed a multiple lists data in an attempt to estimate the number of infants born with a specific congenital anomaly in Massachusetts during a fixed time interval. Five distinct types of sources were considered: obstetric records (183 cases); other hospital records (215 cases); list maintained by state Department of Public Health (36 cases); list maintained by state Department of Mental Health (263 cases), and records by special schools (252 cases). The total number of cases identified was 537. Previous studies [5–7] have shown that inferences regarding the number of missing cases can be made under proper assumptions.

These four data sets will be discussed in Section 5 as illustrations. As indicated by the International Working Group of Disease Monitoring and Forecasting (IWGDMF) [8, 9], similar examples arise in various disease categories such as cancer, drug use, infectious diseases, injuries and others.

The adjustment of undercount has an analogue in the biological sciences: estimating the number of unseen animals in a closed population. Here a closed population means that there is no birth, death or migration so that the population size is a constant during the study period. The estimation of population size is a classical problem and has been extensively discussed in the literature. Although doing complete census counts for some clustered animal populations, especially in colonies, is not impossible, biologists have long realized that complete enumeration is nearly an unattainable ideal for most mobile populations and that proper adjustment for undercount is needed.

Since most animals cannot be drawn like balls in an urn or numbers on a list, traditional random sampling techniques are not easily applicable to biological surveys. To tackle the above undercount problem, special types of sampling schemes have been developed.

Capture-recapture sampling has been widely used to adjust for undercount in the biological sciences [10–13]. It would be unnecessary and almost impossible to count every animal in a closed population in order to obtain an accurate estimate of population size. The recapture information (that is, source-overlap information or source intersection) collected by marking or tagging can be used to estimate the number of missing under proper assumptions.

In contrast, epidemiologists have attempted to enumerate all relevant cases to obtain the prevalence rates for various diseases. Some studies based on health records did ascertain almost all patients [14]. However, as Hook and Regal [15] indicated, most prevalence surveys merging several records of lists are likely to miss some cases and thus be biased to underestimates. There is relatively little literature in the health sciences on the assessment of the completeness of these types of surveys or on the adjustment for under-ascertainment. Therefore, as commented by LaPorte *et al.* [16], people know more about the number of animals than the count of diseases. In this tutorial, we focus on the applications of the closed capture-recapture methods to such epidemiological studies. In the same way that ecologists and biologists count animals, we introduce in this paper the use of capture-recapture models to count human populations.

In Section 2, the capture-recapture technique and its adaptation for use in human populations are reviewed. In Section 3, the general data structure and the concept of two types of dependencies are formulated. Section 4 introduces three closed capture-recapture models which are applicable to ascertainment data in epidemiology. A program developed by the authors is demonstrated to estimate population size in Section 5 using the four data sets described in the beginning of this section. Some remarks about the limitations of capture-recapture methods are discussed in Section 6.

2. CAPTURE-RECAPTURE

2.1. Animal populations

In an animal capture-recapture experiment, traps or nets are placed in the study area and the population is sampled several times. At the first trapping sample a number of animals are captured; the animals are uniquely tagged or marked and released into the population. Then at each subsequent trapping sample we record and attach a unique tag to every unmarked, record the capture of any animal that has been previously tagged, and return all animals to the population. At the end of the experiment the complete capture history for each animal is known. Such experiments are also called mark-recapture, tag-recapture and multiple-record system.

According to Seber [10], the first use of the capture-recapture technique can be traced back to Laplace, who used it to estimate the population size of France in 1786. The earliest applications to ecology include Petersen's and Dahl's work on fish populations and Lincoln's use of band returns to estimate waterfowl in 1930. More sophisticated statistical theory and inference procedures have been proposed since Darroch's paper [17], in which the mathematical framework of this topic was founded. Seber [10–12] and Schwarz and Seber [13] provided comprehensive reviews on the methodology and applications.

The models are generally classified as either closed or open. As stated in Section 1, the size of a closed population is assumed to be a constant throughout the study period. The closure

assumption is usually valid for data collected in a relatively short time during a non-breeding season. In an open model, births and deaths are allowed and thus the size varies with time. Open models are usually used for modelling the data from long-term investigations of animals or migratory birds. The capture-recapture technique has also been adopted to estimate survival rates, birth rates and migration rates.

We restrict this paper to closed models because the population size for most epidemiological studies can be assumed to be approximately a constant in a fixed time period. We also assume that individuals do not lose their marks and all marks are recorded and matched correctly. Marking or tagging is mainly used to distinguish individuals, and thus the recapture information can be applied to evaluate the degree of undercount. When recaptures in the subsequent samples are few, we know intuitively for independent samples that the size is much higher than the number of distinct captures. On the other hand, if the recapture rate is high, then we are likely to have caught most of the animals.

We remark that overlap information in some studies can be obtained without marking. For example, when the main purpose in regular bird counts is to count the number of species, marking or tagging is not needed because species identification for each sighting would suffice for providing species overlap and that information can be properly used to estimate the number of undiscovered species.

2.2. Human populations

The capture-recapture technique originally developed for animal studies has been applied to human populations under the term 'multiple-record system' [6, 8, 9, 18–21]. The special two-sample cases are often referred to as the 'dual-system' or 'dual-record system'. For ascertainment data, if each list is regarded as a trapping sample and identification numbers and/or names are used as tags or marks, then this framework is similar to a capture-recapture set-up for wildlife estimation. Thus 'capture in a sample' corresponds to 'being recorded or identified in a list', and 'capture probability' becomes 'ascertainment probability'.

The earliest references to the application of the capture-recapture techniques to health science included the pioneering paper by Sekar and Deming [18] for two samples, Wittes and Sidel [19] for three samples, Wittes [20] for four samples, Wittes *et al.* [5] and Fienberg [6] for five samples. Epidemiologists recently have shown renewed and growing interest in the use of the capture-recapture models [16, 21]. Hook and Regal [22] also suggested the use of capture-recapture models even for apparently exhaustive surveys. Hook and Regal [23], IWGDMF [8, 9] and Chao [24] provided overviews of the applications of the capture-recapture models specifically to epidemiological data. However, some critical comments and practical concerns about the use of capture-recapture models have been expressed by several authors [14, 25–28]. We will address their main concerns in Section 6.

Three main differences between wildlife and human applications are noted:

(i) There are usually more trapping samples in wildlife studies, whereas in most epidemiological surveys only two to four lists are available.
(ii) There is a natural time ordering in animal experiments, but generally no such order exists in epidemiological lists, or the order may vary with individuals.
(iii) In animal studies, identical trapping methods are usually used in all trapping samples. Hence animals' behavioural response to capture is often present and is modelled in analysis. In human populations, different types of ascertainment sources are utilized to

Table I. Individual ascertainment data for three lists.

Individual	List 1	List 2	List 3
1	X_{11}	X_{12}	X_{13}
2	X_{21}	X_{22}	X_{23}
....	...	...	...
M	X_{M1}	X_{M2}	X_{M3}
$M+1$	0	0	0
....	...	...	...
N	0	0	0

search all individuals. The behavioural response due to the sampling scheme is not commonly considered in models.

Researchers in wildlife and human populations have developed models and methodologies along separate lines. Three of these approaches will be introduced after the formulation of the data structure and the concept of dependence among samples.

3. DATA STRUCTURE AND DEPENDENCE

3.1. Data structure

We first introduce some notation. Assume that the true population size is N which is our parameter of interest. The individuals can be conceptually indexed by $1,2,\ldots,N$ and all individuals act independently. Assume that there are t samples (lists, records or sources) and they are indexed by $1,2,\ldots,t$. Presence and absence in any source are denoted by 1 and 0, respectively. For a three-list case as given in Table I, we can use three numbers (each is either 0 or 1) to denote the record of each individual. For example, individual 1 was identified in list 1 only. Then it is associated with the record (100) in the data; individual 2 was identified in all three lists, it is recorded as the record (111). Each individual in a three-list case is associated with one of the following seven possible 'capture histories' or 'ascertainment records': (001); (010); (011); (100); (101); (110), and (111). Suppose there are M identified individuals and $N-M$ uncounted. Without losing generality, assume that these M identified individuals are indexed by $1,2,\ldots,M$. If we augment all the identified records by $N-M$ individuals with history (000) as in Table I, then the ascertainment data for all individuals can be conveniently expressed by an $N \times t$ matrix $X=(X_{ij})$. Here $X_{ij}=1$ if the ith individual is listed in the jth sample, 0 otherwise. A record (000) means that an individual is not identified in any of the three samples.

The ascertainment data for all identified individuals can be aggregated as a categorical data format as shown in Table II for the HAV data. That is, the frequencies of the same record are grouped. Let $Z_{s_1 s_2 \ldots s_t}$ be the number of individuals with record $s_1, s_2, \ldots, s_t$, where $s_j=0$ denotes absence in sample j and $s_j=1$ denotes presence in sample j. For example, when $t=3$, there are seven observed cells Z_{001}, Z_{010}, Z_{011}, Z_{100}, Z_{101}, Z_{110} and Z_{111}, where Z_{001} is the number of individuals listed in sample 3 only, Z_{011} is the number of individuals listed in samples 2 and 3 but not in sample 1. A similar interpretation pertains

Table II. Aggregated data on hepatitis A virus.

Hepatitis A list			Data
P	Q	E	
0	0	0	$Z_{000} = ??$
0	0	1	$Z_{001} = 63$
0	1	0	$Z_{010} = 55$
0	1	1	$Z_{011} = 18$
1	0	0	$Z_{100} = 69$
1	0	1	$Z_{101} = 17$
1	1	0	$Z_{110} = 21$
1	1	1	$Z_{111} = 28$

to other capture histories. The missing cell $Z_{000} = N - M$ denotes the uncounted. When we add over a sample, the subscript corresponding to that sample is replaced by a '+' sign. For example, $Z_{+11} = Z_{011} + Z_{111}$ and $Z_{++1} = Z_{001} + Z_{011} + Z_{101} + Z_{111}$, and $Z_{+++} = N$. Let n_j, $j = 1,2,\ldots,t$ be the number of individuals listed in sample j. For $t = 3$, we have $n_1 = Z_{1++}$, $n_2 = Z_{+1+}$, $n_3 = Z_{++1}$.

For the HAV data in Table II, there were 63 people listed in the E-list only, 55 people listed in the Q-list only, and 18 people listed in both lists Q and E but not in the P-list. Similarly, we can interpret the other records. The purpose here is to estimate the number of total individuals (that is, N) who were infected in the outbreak. It is thus equivalent to predicting the number of missed (that is, $Z_{000} = N - M$) by all three sources.

In a typical approach in epidemiology, cases in various lists are merged and any duplicate cases are eliminated. That is, the capture histories in Table I and the categories in Table II are ignored in the analysis and only the final merged list is obtained. This typical approach assumes complete ascertainment and does not correct or adjust for under-ascertainment. However, there were non-negligible uncounted cases in many epidemiological surveillance studies. For example, before the screen serological check for all students of that college, epidemiologists suspected that the observed number of cases (271) in Table II considerably undercounted the true number of infected and an evaluation of the degree of undercount was needed [1, 29].

3.2. Dependence among samples

A crucial assumption in the traditional statistical approach is that the samples are independent. Since individuals can be cross-classified according to their presence or absence in each list, the dependence for any two samples is usually interpreted from a 2×2 categorical data analysis in human applications. In animal studies, traditional 'equal-catchability assumption' is even more restrictive, that is, in each fixed sample all animals including marked and unmarked have the same capture probability. (Equal catchability assumption implies independence among samples but the reverse is not true; see Section 4.3.) Non-independence or unequal catchabilities may be caused by the following two sources:

(i) Local dependence (also called list dependence or local list dependence) within each individual; conditional on any individual, the inclusion in one source has a direct causal

effect on his/her inclusion in other sources. That is, the response of a selected individual to one source depends on his/her response to the other sources. For example, the probability of going to a hospital for treatment for any individual depends on his/her result on the serum test of the HAV. The ascertainment of the serum sample and that of the hospital sample then becomes dependent. We remark that 'local independence' has been a fundamental assumption in many statistical methodologies [30].

(ii) Heterogeneity between individuals; even if the two lists are independent within individuals, the ascertainment of the two sources may become dependent if the capture probabilities are heterogeneous among individuals. This phenomenon is similar to Simpson's paradox in categorical data analysis. That is, aggregating two independent 2×2 tables might result in a dependent table. Hook and Regal [31] presented an interesting epidemiological example.

These two types of dependencies are usually confounded and cannot be easily disentangled in a data analysis. Lack of independence leads to a bias (called 'correlation bias' in census undercount estimation [32]) for the usual estimator which assumes independence. We use a two-sample animal experiment to explain the direction of the bias. Assume that a first sample of n_1 animals is captured. Therefore, the marked rate in the population is n_1/N. A second sample of n_2 animals is subsequently drawn and there are m_2 (that is, Z_{11} in our notation for grouped data) previously marked. The capture rate for the marked (recapture rate, overlap rate) in the second sample can be estimated by m_2/n_2. If the two samples are independent, then the recapture rate should be approximately equal to the marked rate in the population. Therefore we have $m_2/n_2 = n_1/N$, which yields an estimate of the population size under independence: $\hat{N}_P = n_1 n_2/m_2$ (the well-known Petersen estimator or dual-system estimator). However, if the two samples are positively correlated, then those individuals captured in the first sample are more easily captured in the second sample. The recapture rate in the second sample tends to be larger than the marked rate in the population. That is, we would expect that $m_2/n_2 > n_1/N$, which yields $N > n_1 n_2/m_2$. As a result, Petersen's estimator underestimates the true size if both samples are positively dependent. Conversely, it overestimates for negatively dependent samples. A similar argument is also valid for a general number of samples. That is, a higher (lower) overlap rate is observed for positively (negatively) dependent samples, which implies fewer (more) estimated missing cases. Therefore, a negative (positive) bias exists for any estimator which assumes independence.

When only two lists are available, three cells are observable: people identified in list 1 only; people identified in list 2 only, and people listed in both. However, there are four parameters: N, two mean capture probabilities and a dependence measure. The data are insufficient for estimating dependence unless additional covariates are available. All existing methods unavoidably encounter this problem and adopt the independence assumption. This independence assumption has become the main weak point in the use of the capture-recapture method for two lists.

A variety of models incorporating dependence among samples have been proposed in the literature. We will review three classes of models: ecological models; log-linear models, and the sample coverage approach. The latter two approaches can be used to provide estimates for some ecological models, but they are considered separately because of their different ways of dealing with dependence.

Table III. Two types of capture probabilities for ecological models.

Model	Multiplicative model in log-linear form	Logistic model
$\mathbf{M}_{tbh}$	$\log(P_{ij}) = \alpha_i + \beta_j + \gamma Y_{ij}$	$\text{logit}(P_{ij}) = \alpha_i + \beta_j + \gamma Y_{ij}$
$\mathbf{M}_{bh}$	$\log(P_{ij}) = \alpha_i + \gamma Y_{ij}$	$\text{logit}(P_{ij}) = \alpha_i + \gamma Y_{ij}$
$\mathbf{M}_{tb}$	$\log(P_{ij}) = \beta_j + \gamma Y_{ij}$	$\text{logit}(P_{ij}) = \beta_j + \gamma Y_{ij}$
$\mathbf{M}_{th}$	$\log(P_{ij}) = \alpha_i + \beta_j$	$\text{logit}(P_{ij}) = \alpha_i + \beta_j$ (Rasch model)
$\mathbf{M}_{h}$	$\log(P_{ij}) = \alpha_i$	$\text{logit}(P_{ij}) = \alpha_i$
$\mathbf{M}_{b}$	$\log(P_{ij}) = \alpha + \gamma Y_{ij}$ $(\alpha_i \equiv \alpha)$	$\text{logit}(P_{ij}) = \alpha + \gamma Y_{ij}$ $(\alpha_i \equiv \alpha)$
$\mathbf{M}_{t}$	$\log(P_{ij}) = \beta_j$	$\text{logit}(P_{ij}) = \beta_j$

4. MODELS AND ESTIMATORS

4.1. Ecological models

This approach specifies various forms of capture probabilities based on empirical investigations of animal ecology. Although most authors in this field did not aim to model dependence between samples, dependence is induced when some special types of capture probabilities are formulated. Two types of probabilities have been proposed: multiplicative and logistic.

The multiplicative class of models was first proposed by Pollock [33] and was fully discussed in the two monographs by Otis *et al.* [34] and White *et al.* [35]. Three sources of variation in capture probability are considered: time-varying, behavioural response, and heterogeneity. The corresponding models are denoted by model $\mathbf{M}_t$, $\mathbf{M}_b$ and $\mathbf{M}_h$, respectively. Various combinations of these three types of unequal capture probabilities (that is, models $\mathbf{M}_{tb}$, $\mathbf{M}_{th}$, $\mathbf{M}_{bh}$ and $\mathbf{M}_{tbh}$) are also considered. These models specify the conditional probability of capturing the ith animal in the jth sample given the capture history of samples $1,2,\ldots,j-1$. Denote this conditional probability by P_{ij} for notational simplicity. A multiplicative form of model $\mathbf{M}_{tbh}$ is

$$P_{ij} = \begin{cases} p_i e_j & \text{until first capture} \\ \phi p_i e_j & \text{for any recapture} \end{cases}$$

where $0 < p_i e_j,\ \phi p_i e_j < 1$. Here the parameters $\{e_1, e_2, \ldots, e_t\}$, $\{p_1, p_2, \ldots, p_N\}$ and ϕ are used to model the time effects, individual heterogeneity and the behavioural response to capture, respectively. Reparameterize $\alpha_i = \log(p_i)$, $\beta_j = \log(e_j)$, $\gamma = \log(\phi)$, and define $Y_{ij} = I$ [the ith animal has been captured before the jth sample] where $I(A)$ is an indicator function of the event A, that is, $I(A) = 1$ if A is true and $I(A) = 0$ otherwise. The time-dependent variable Y_{ij} is used to denote the prior capture history of individual i for sample j. Then the multiplicative type of probability can be conveniently expressed as the following log-linear form:

$$\log(P_{ij}) = \alpha_i + \beta_j + \gamma Y_{ij} \tag{1}$$

All submodels can be easily formulated as shown in Table III.

Logistic types of models have also been proposed [36–38] in the literature and the form of a logistic model $\mathbf{M}_{tbh}$ is

$$\text{logit}(P_{ij}) \equiv \log\left(\frac{P_{ij}}{1 - P_{ij}}\right) = \alpha_i + \beta_j + \gamma Y_{ij} \tag{2}$$

which is equivalent to $P_{ij} = \exp(\alpha_i + \beta_j + \gamma Y_{ij})/[1 + \exp(\alpha_i + \beta_j + \gamma Y_{ij})]$. All submodels of the logistic form are also shown in Table III. The two types of models and submodels can thus be integrated in a unified expression and they only differ in the link function. The logistic model $\mathbf{M}_{th}$, that is, $\gamma = 0$ in equation (2), is the well-known Rasch model [39], which plays an important role in educational statistics and in the analysis of survey data.

To reduce the number of parameters and to remove the non-identification caused by the numerous parameters $\{\alpha_1, \alpha_2, \ldots, \alpha_N\}$, one of the following is usually assumed: (i) they are a random sample from a parametric distribution characterized by a few parameters, for example, beta or normal distributions (random-effects model) [40, 41]; (ii) they can be partitioned as two or more homogeneous groups (latent class model or mixture model) [42]; (iii) They are determined by the first two moments totally (usually, mean and coefficient of variation) [43].

For both types of ecological models, the samples are independent only for model $\mathbf{M}_t$. Local list dependence is present for models $\mathbf{M}_b$ and $\mathbf{M}_{tb}$; heterogeneity arises for model $\mathbf{M}_h$; both types of dependencies exist for models $\mathbf{M}_{bh}$ and $\mathbf{M}_{tbh}$. As noted in the reviews [10–13], estimators for various models may be found in the literature. These estimators rely on a wide range of methodologies.

An analysis of the closed model can be performed using a comprehensive computer program, CAPTURE [44]. The program is readily available from Gary White's website (http://www.cnr.colostate.edu/~gwhite/software.html). See Section 5 for another program calculating some recent estimates and their standard errors.

As indicated in Section 2.2, there is a natural time ordering in ecological experiments. Models with behavioural response (that is, models $\mathbf{M}_b$, $\mathbf{M}_{tb}$, $\mathbf{M}_{tb}$ and $\mathbf{M}_{tbh}$) allow the capture of any individual depending on its own 'previous' capture histories and thus ordering is implicitly involved in these four models. Meanwhile, almost all estimation procedures derived under these models depend on the ordering of the lists. Since there is usually no sequential time order in ascertainment lists or sources, any model involving behavioural response or any estimator depending on the list order has limited use in epidemiology. Therefore, only models $\mathbf{M}_t$, $\mathbf{M}_h$ and $\mathbf{M}_{th}$ are potentially useful for our applications.

For model $\mathbf{M}_t$, the multiplicative model and the logistic model are equivalent because either model is only a reparameterization of the other. The MLE (conditional on the total number of identified) of population size under model $\mathbf{M}_t$ is identical to that obtained under the independent log-linear model [10, 17, 45]. Therefore, model $\mathbf{M}_t$ will be included in the log-linear model approach in Section 4.2.

For model $\mathbf{M}_h$, both types of models are also equivalent. Model $\mathbf{M}_h$ assumes that each individual has its own unique probability that remains constant over samples. Thus it is usually applied to the situations where similar types of trapping methods are taken. Model $\mathbf{M}_{th}$ extends model $\mathbf{M}_h$ by allowing for various sampling efforts or time effects. A widely used estimator for model $\mathbf{M}_h$ is the jackknife estimator proposed by Burnham and Overton [46]. The jackknife estimator is a linear function of the capture frequencies $\{f_1, f_2, \ldots, f_t\}$ where f_k denotes the number of animals captured exactly k times in the t samples. The jackknife estimator is invariant to the order of the lists. As indicated by Otis *et al.* in their monograph (reference [34], p. 34), the bias of the jackknife is within a tolerable range if the number of trapping samples is greater than five.

The logistic model $\mathbf{M}_{th}$ (that is, the Rasch model) is equivalent to a quasi-symmetric log-linear model with some moment constraints [32, 47]. Hence it will be discussed in Section 4.2. For multiplicative models $\mathbf{M}_{th}$ and $\mathbf{M}_h$, Chao *et al.* [48] and Lee and Chao [43] proposed

some estimators focusing on animal data, but those estimators are suggested for use when the number of samples is sufficiently large (say at least five, as in the jack-knife method). Therefore, except for the Rasch model, heterogeneous ecological models are recommended only when at least five lists are available. An example with five samples is given in Section 5.4 for illustration.

4.2. Log-linear models

The log-linear models have been proposed [6, 40, 45, 47, 49, 50] to handle dependence among samples. This approach is well discussed in the two review papers by IWGDMF [8, 9], thus we only provide a brief description here. In this approach, the data are regarded as a form of an incomplete 2^t contingency table (t is the number of lists) for which the cell corresponding to those individuals unlisted by all samples is missing. Then various log-linear models are fitted to the observed cells. How well a model fits the data is assessed using the deviance statistic and a model is usually selected based on the Akaiki information criterion. The chosen model is then projected onto the unobserved cell by assuming that there is no t-sample interaction. The two types of dependencies can be modelled by including some specific interactions or common interaction in the models.

We use the three-list data for illustration. The log-linear approach models the logarithm of the expected value of each observable category, that is, the most general model is

$$\log E(Z_{ijk}) = u + u_1 I(i=1) + u_2 I(j=1) + u_3 I(k=1) + u_{12} I(i=j=1) + u_{13} I(i=k=1) + u_{23} I(j=k=1) + u_{123} I(i=j=k=1) \tag{3}$$

That is, $\log E(Z_{001}) = u + u_3$, $\log E(Z_{010}) = u + u_2$, $\log E(Z_{100}) = u + u_1$, $\log E(Z_{110}) = u + u_1 + u_2 + u_{12}$, $\log E(Z_{101}) = u + u_1 + u_3 + u_{13}$, $\log E(Z_{011}) = u + u_2 + u_3 + u_{23}$, and $\log E(Z_{111}) = u + u_1 + u_2 + u_3 + u_{12} + u_{13} + u_{23} + u_{123}$. This is a reparameterization of the eight expected values.

For three-list data, we have seven observed categories, whereas there are eight parameters in equation (3). Therefore, a natural assumption is that there is no three-list interaction term, that is, $u_{123} = 0$. Intuitively, this means the complete 2×2 table formed with respect to lists 2 and 3 for individuals in list 1 and the incomplete 2×2 table for individuals not in list 1 have the same odds ratio. The sample odds ratio for the former table is $Z_{111}Z_{100}/(Z_{110}Z_{101})$ whereas the odds ratio for the latter table is $Z_{011}Z_{000}/(Z_{010}Z_{001})$. The assumption of $u_{123} = 0$ allows the following extrapolation formula:

$$\hat{Z}_{000} = \hat{Z}_{001}\hat{Z}_{010}\hat{Z}_{100}\hat{Z}_{111}/(\hat{Z}_{110}\hat{Z}_{011}\hat{Z}_{101}) \tag{4}$$

which expresses the estimated missing cases as a function of the fitted values of other categories [6, 45]. The fitted values of the observable cells are determined by the chosen model.

The independent model includes only main effects as given by

$$\log E(Z_{ijk}) = u + u_1 I(i=1) + u_2 I(j=1) + u_3 I(k=1)$$

The resulting estimator under this model using (4) is equivalent to the MLE for model $\mathbf{M}_t$ [45]. The interaction terms are used to model dependence. If local list dependence arises in samples 1 and 2, then the interaction term u_{12} is included, and the model is denoted as model

(12, 3) or 12/3 as used in categorical data analysis. If local dependence also appears in samples 1 and 3, then the two interactions u_{12} and u_{13} are needed. The model is denoted as model (12, 13) or 12/13 and similarly for models 13/2, 23/1, 13/23 and others.

The log-linear model can also be motivated by the Rasch model and its generalizations which incorporate heterogeneity among individuals. As shown in Table III and Section 4.1, the Rasch model assumes $\text{logit}(P_{ij}) = \alpha_i + \beta_j$. Only dependence due to heterogeneity arises in this model and there is no local list dependence. A generalized Rasch model allows the heterogeneity effects $\{\alpha_1, \alpha_2, \ldots, \alpha_N\}$ to be different for two or more separate groups of samples. For example, in a three-list case, it assumes

$$\text{logit}(P_{ij}) = \begin{cases} \alpha_i + \beta_j, & j = 1,2 \\ \alpha_i^* + \beta_j, & j = 3 \end{cases} \tag{5}$$

It has been verified [32, 47, 51] that the Rasch (generalized Rasch) model is equivalent to a quasi-symmetric (partial quasi-symmetric) model with some moment constraints. Except for the constraints, a quasi-symmetric model for the three-list case with no second-order interaction, that is, $u_{123} = 0$, is equivalent to the model with first-order interactions identical; this is denoted by $(12 = 13 = 23)$ or simply H1 (first-order heterogeneity [8, 9]). Only one degree of freedom is used to model heterogeneity. A partial quasi-symmetric model in equation (5) with $u_{123} = 0$ is equivalent to the model with $u_{13} = u_{23}$. This model is denoted as $(13 = 23, 12)$. Similarly, we have models $(12 = 13, 23)$ and $(12 = 23, 13)$ corresponding to other two partial quasi-symmetric models. Therefore, the dependence due to heterogeneity can be modelled by either a quasi-symmetric or a partial quasi-symmetric model. We remark that when both types of dependencies occur, they are inevitably confounded in the interaction or common interaction terms and cannot be separated.

The model in (3) can be similarly formulated when there are more than three lists. The basic assumption for four lists is the third-order interaction vanishes (that is, $u_{1234} = 0$); that is, the three-list interaction for individuals in list 1 is the same as that for individuals not in list 1. Local list dependence can be modelled by including the first-order interaction term $(u_{12}, u_{13}, u_{14}, u_{23}, u_{24}, u_{34})$ and/or the second-order interaction $(u_{123}, u_{134}, u_{124}, u_{234})$. The Rasch model is equivalent to a model with first-order heterogeneity H1 (that is, $12 = 13 = 14 = 23 = 24 = 34$) and second-order heterogeneity H2 (that is, $123 = 124 = 134 = 234$). Thus two parameters are used to model heterogeneity in the Rasch model for four lists. If additional local dependencies also occur between lists 1 and 2, lists 1 and 3, and lists 2 and 4, then we add three more parameters u_{12}, u_{13} and u_{24} to the model and the resulting model is denoted as (12/13/24, H1, H2). Refer to the sample program output in Section 5.3 for other types of models. For the general case of t lists, the Rasch model with no highest order interaction is equivalent to a model with interactions of equal order assumed identical; a partial quasi-symmetric Rasch model is equivalent to assuming some of the interactions are equal; see Lloyd [51] for details.

Other useful models, including normal random-effects models, latent class models [40, 42, 50], non-parametric models [52] and Bayesian approaches [47], have also been proposed to reflect heterogeneity in the Rasch model, thus widening the scope and application areas of the log-linear model approach.

4.3. Sample coverage approach

As discussed in Section 2.1, overlap information plays an important role in estimating the number of missing cases. The main purposes of the sample coverage approach proposed by Chao and Tsay [53] and Tsay and Chao [54] are to provide a measure to quantify the overlap information and also to propose parameters to quantify source dependence.

Dependence is modelled by parameters called the 'coefficient of covariation' (CCV). To better understand the CCV parameters, we only consider heterogeneity. Define P_{ij} as the conditional probability of identifying individual i in the jth list. The CCV of samples j and k for a fixed-effect approach is defined as

$$\gamma_{jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{(P_{ij}-\mu_j)(P_{ik}-\mu_k)}{\mu_j\mu_k} = \frac{1}{N}\frac{\sum_{i=1}^{N}P_{ij}P_{ik}}{\mu_j\mu_k} - 1 \tag{6a}$$

where $\mu_j = \sum_{i=1}^{N} P_{ij}/N = E(n_j)/N$ denotes the average probability for being listed in the jth sample. The magnitude of γ_{jk} measures the degree of dependence between samples j and k. The two heterogeneous samples are independent if and only if $\gamma_{jk}=0$, that is, $N^{-1}\sum_{i=1}^{N}P_{ij}P_{ik} = \mu_j\mu_k$, which means that the covariance between the two sets of probabilities, $\{P_{ij};\ i=1,2,\ldots,N\}$ and $\{P_{ik};\ i=1,2,\ldots,N\}$, is zero. They are positively (negatively) dependent if $\gamma_{jk}>0$ ($\gamma_{jk}<0$), which is equivalent to $N^{-1}\sum_{i=1}^{N}P_{ij}P_{ik} > \mu_j\mu_k$ ($N^{-1}\sum_{i=1}^{N}P_{ij}P_{ik} < \mu_j\mu_k$), that is, the average probability of jointly being listed in the two samples is greater (less) than that in the independent case. It follows from (6a) that the CCV is zero if the capture probabilities for one sample are constant (that is, a random sample); in this case, no correlation bias arises even if the other sample is highly heterogeneous provided that no local dependence exists. We can prevent local dependence by taking the second sample random because equal catchability implies probabilities are the same regardless of marking status. For a random-effect model, we assume that $\{(P_{i1},P_{i2},\ldots,P_{it}),\ i=1,2,\ldots,N\}$ are a random sample from a t-dimensional distribution $F_{P_1,P_2,\ldots,P_t}(p_1,p_2,\ldots,p_t)$. The CCV for samples j and k then becomes

$$\gamma_{jk} = \frac{E[(P_j-\mu_j)(P_k-\mu_k)]}{\mu_j\mu_k} = \frac{\mathrm{cov}(P_j,P_k)}{\mu_j\mu_k} = \frac{E(P_jP_k)}{\mu_j\mu_k} - 1 \tag{6b}$$

where $\mu_j = E(P_j)$ denotes the average capture probability for the jth sample. Researchers in fishery sciences have suggested that correlation bias due to heterogeneity could be reduced if two different sampling schemes were used (for example, trapping and then resighting, or netting and then angling). This was justified by Seber (reference [10], p. 86); it also could be seen from formula (6b) because there is almost no covariance between the distributions for two distinct samplings.

The CCV for more than two samples can be similarly defined and interpreted. For example, the CCV for samples $k_1,k_2,\ldots,k_m$ in a random-effect model is defined as

$$\gamma_{k_1k_2\ldots k_m} = \frac{E[(P_{k_1}-\mu_{k_1})(P_{k_2}-\mu_{k_2})\ldots(P_{k_m}-\mu_{k_m})]}{\mu_{k_1}\mu_{k_2}\ldots\mu_{k_m}}$$

The CCV for the general cases with two types of dependencies has been developed [53], but it will not be addressed here. We only remark that all CCVs in the general cases measure the overall effect of the two types of dependencies.

For two lists, the usual independence assumption is equivalent to setting the two-sample CCV at 0 ($\gamma_{12}=0$). It is not possible to model dependence between two lists as we discussed

in Section 3.2. For the three-list cases, there are seven observable categories (as in the HAV data in Table II) and eight parameters: $N; \mu_1; \mu_2; \mu_3; \gamma_{12}; \gamma_{13}; \gamma_{23}; \gamma_{123}$. One constraint is still needed, yet it is possible to model dependence. Consequently, at least three samples are required to reasonably estimate any dependence parameters.

The concept of sample coverage was originally proposed by Turing and Good [55]. This concept has played an important role in the classical species estimation for heterogeneous communities [56] and has been modified for multiple-sample cases [53], in which the sample coverage is used as a measure of overlap fraction. The basic idea is that the sample coverage can be well estimated even in the presence of two types of dependencies. Thus an estimate of population size can be derived via the relationship between the population size and the sample coverage. Chao and Tsay [53] dealt mainly with the three-sample case. Extension to cases with more than three samples was provided by Tsay and Chao [54]. Below we will separately summarize the estimation procedures for the three-list case and the general case.

If an additional case were selected from the third list, then a proper overlapping measure would be the conditional probability of finding this case that had already been identified in the combined list of the other two sources (that is, finding a case i for which $X_{i1} + X_{i2} > 0$). The overlap fraction can be quantified as $\sum_{i=1}^{N} P_{i3} I(X_{i1} + X_{i2} > 0) / \sum_{i=1}^{N} P_{i3}$. Considering that this additional individual could be selected from any of the three lists, we define the sample coverage as the average of the three possible overlap fractions as follows:

$$C = \frac{1}{3}\left[\frac{\sum_{i=1}^{N} P_{i3} I(X_{i1} + X_{i2} > 0)}{\sum_{i=1}^{N} P_{i3}} + \frac{\sum_{i=1}^{N} P_{i2} I(X_{i1} + X_{i3} > 0)}{\sum_{i=1}^{N} P_{i2}} + \frac{\sum_{i=1}^{N} P_{i1} I(X_{i2} + X_{i3} > 0)}{\sum_{i=1}^{N} P_{i1}}\right]$$

An estimator of the sample coverage is [53]

$$\hat{C} = 1 - \frac{1}{3}\left(\frac{Z_{100}}{n_1} + \frac{Z_{010}}{n_2} + \frac{Z_{001}}{n_3}\right) = \frac{1}{3}\left[\left(1 - \frac{Z_{100}}{n_1}\right) + \left(1 - \frac{Z_{010}}{n_2}\right) + \left(1 - \frac{Z_{001}}{n_3}\right)\right] \tag{7}$$

which is the average (over three lists) of the fraction of cases found more than once. Note that Z_{100}, Z_{010} and Z_{001} are the numbers of individuals listed only in one sample. Hence this estimator is the complement of the fraction of singletons. Obviously, singletons cannot contain any overlapping information. Define

$$D = \frac{1}{3}[(M - Z_{100}) + (M - Z_{010}) + (M - Z_{001})] = M - \frac{1}{3}(Z_{100} + Z_{010} + Z_{001}) \tag{8}$$

Here $(Z_{100} + Z_{100} + Z_{100})/3$ represents the average of the non-overlapped cases and recall that M denotes the total number of identified cases. Thus D can be interpreted as the average of the overlapped cases. The sample coverage estimation procedures for the three-list case are summarized in the following:

1. When the three sources are independent, a simple population size estimator is derived as

$$\hat{N}_0 = D/\hat{C} \tag{9}$$

The above estimator is obtained by noting that C is reduced to D/N under independence. It can also be intuitively thought of as ratio of overlapped cases to overlap fraction.

2. When dependence exists and the overlap information is large enough (how large it should be will be discussed further below), we take into account the dependence by adjusting the simple estimator given in (9) based on a function of two-sample CCVs. The adjustment expansion formula [54] is

$$N = \frac{D}{\hat{C}} + \frac{1}{3\hat{C}}[(Z_{1+0} + Z_{+10})\gamma_{12} + (Z_{10+} + Z_{+01})\gamma_{13} + (Z_{01+} + Z_{0+1})\gamma_{23}] + R \tag{10}$$

where R is the remainder term in the above expansion and $R/N \to \psi = \mu_1\mu_2[\gamma_{12}(\gamma_{13}+\gamma_{23}) - \gamma_{123}] + \mu_1\mu_3[\gamma_{13}(\gamma_{12}+\gamma_{23}) - \gamma_{123}] + \mu_2\mu_3[\gamma_{23}(\gamma_{12}+\gamma_{13}) - \gamma_{123}]$ in probability when N becomes large. Our constraint is set by letting $\psi = 0$. The constraint is satisfied under a multiplicative model $\mathbf{M}_{\text{th}}$ where the heterogeneity effects follow a gamma type of distribution [53]. Since gamma distribution can cover a wide range of heterogeneity patterns, this is our main motivation for setting this constraint. In equation (10), if R is ignored and CCVs are substituted by the following functions of N:

$$\gamma_{13} = N\frac{Z_{1+1}}{n_1 n_3} - 1, \quad \gamma_{23} = N\frac{Z_{+11}}{n_2 n_3} - 1, \quad \gamma_{12} = N\frac{Z_{11+}}{n_1 n_2} - 1 \tag{11}$$

then we end up with an estimating equation of N. The following solution of the resulting estimating equation is the estimator:

$$\hat{N} = \left[\frac{Z_{+11} + Z_{1+1} + Z_{11+}}{3\hat{C}}\right] \Bigg/ \left\{1 - \frac{1}{3\hat{C}}\left[\frac{(Z_{1+0} + Z_{+10})Z_{11+}}{n_1 n_2} + \frac{(Z_{10+} + Z_{+01})Z_{1+1}}{n_1 n_3} + \frac{(Z_{0+1} + Z_{01+})Z_{+11}}{n_2 n_3}\right]\right\} \tag{12}$$

3. For relatively low sample coverage data, we feel the data do not contain sufficient information to accurately estimate the population size. In this case, the following 'one-step' estimator $\hat{N}_1$ is suggested (the estimator is called 'one-step' because it is obtained by one iterative step from the adjustment formula (10) with γ_{ij}'s being replaced by (11)):

$$\hat{N}_1 = \frac{D}{\hat{C}} + \frac{1}{3\hat{C}}[(Z_{1+0} + Z_{+10})\hat{\gamma}_{12} + (Z_{10+} + Z_{+01})\hat{\gamma}_{13} + (Z_{01+} + Z_{0+1})\hat{\gamma}_{23}] \tag{13}$$

where CCV estimates are

$$\hat{\gamma}_{13} = \hat{N}'\frac{Z_{1+1}}{n_1 n_3} - 1, \quad \hat{\gamma}_{23} = \hat{N}'\frac{Z_{+11}}{n_2 n_3} - 1, \quad \hat{\gamma}_{12} = \hat{N}'\frac{Z_{11+}}{n_1 n_2} - 1 \tag{13a}$$

and

$$\hat{N}' = \frac{D}{\hat{C}} + \frac{1}{3\hat{C}}\left[(Z_{1+0} + Z_{+10})\left(\frac{D}{\hat{C}}\frac{Z_{11+}}{n_1 n_2} - 1\right) + (Z_{10+} + Z_{+01})\left(\frac{D}{\hat{C}}\frac{Z_{1+1}}{n_1 n_3} - 1\right) + (Z_{01+} + Z_{0+1})\left(\frac{D}{\hat{C}}\frac{Z_{+11}}{n_2 n_3} - 1\right)\right]$$

This one-step estimator can be regarded as a lower (upper) bound for positively (negatively) dependent samples. Hook and Regal [7] noted that most data sets used in epidemiological applications tend to have a net positive dependence. Thus the one-step estimator is often used as a lower bound as will be shown in Section 5.

A bootstrap resampling method [57] is proposed to obtain estimated standard errors for the above three estimators and to construct confidence intervals using a log-transformation [58]. A relatively low overlap fraction means that there are relatively many singletons. In this case, the undercount cannot be measured accurately due to insufficient overlap. Consequently, a large standard error is usually associated with the estimator in equation (12). How large should the overlap information be? Previous simulation studies [29] have suggested that the estimated sample coverage should be at least 55 per cent. A more practical data-dependent guideline can be determined from the estimated bootstrap SE associated with the estimator given in (12). If the estimated bootstrap standard error becomes unacceptable (say, it exceeds one-third of the population size estimate), then only the lower or upper bound in (13) is recommended.

We now outline the result for the general t-sample case. The sample coverage for the general case can be defined in a similar manner and the estimator is

$$\hat{C} = 1 - \frac{1}{t}\sum_{k=1}^{t}\frac{S_k}{n_k} \tag{14}$$

where $S_n = Z_{k_1k_2\ldots k_t} I[k_n = 1, k_j = 0, j \neq n]$ denotes the number of individuals that are listed in sample n only (that is, singletons). For $t = 4$, we have $S_1 = Z_{1000}$, $S_2 = Z_{0100}$, $S_3 = Z_{0010}$, and $S_4 = Z_{0001}$ and $\hat{C} = 1 - (Z_{1000}/n_1 + Z_{0100}/n_2 + Z_{0010}/n_3 + Z_{0001}/n_4)$, which is an extension of (7). In the independent case, a valid estimator is $D/\hat{C}$, where

$$D = \frac{1}{t}\sum_{k=1}^{t}\left\{\sum_{i=1}^{N} I\left[\sum_{j\neq k} X_{ij} > 0\right]\right\} = M - \frac{1}{t}\sum_{k=1}^{t} S_k \tag{15}$$

Define $H(i,j) = Z_{k_1k_2\ldots k_t}\, I[k_i = 1, k_j = +, k_n = 0, n \neq i, n \neq j]$, and $A(i,j) = H(i,j) + H(j,i)$. For example, $A(1,2) = Z_{1+00} + Z_{+100}$, $A(2,3) = Z_{01+0} + Z_{0+10}$. When any type of dependence exists, a generalized formula of (10) becomes [54]

$$N = \frac{D}{\hat{C}} + \frac{1}{t\hat{C}}\sum\sum_{i<j} A(i,j)\gamma_{ij} + R^*$$

where R^* denotes the remainder term. As in the three-list case, R^*/N tends to zero under a multiplicative model $\mathbf{M}_{\text{th}}$ where the heterogeneity effects follow a gamma type of distribution when N is large enough. Let $B(i,j) = Z_{k_1k_2\ldots k_t} I[k_i = k_j = 1, k_n = +, n \neq i, n \neq j]$. For example, $B(1,2) = Z_{11++}$ and $B(2,3) = Z_{+11+}$. Using the relationship that $\gamma_{ij} = NB(i,j)/(n_i n_j) - 1$ and substituting it into the above equation, an estimator for the t-sample case can be shown to be

$$\hat{N} = \left[\frac{D}{\hat{C}} - \frac{1}{t\hat{C}}\sum\sum_{i<j} A(i,j)\right]\left\{1 - \frac{1}{t\hat{C}}\sum\sum_{i<j}\frac{A(i,j)B(i,j)}{n_i n_j}\right\}^{-1} \tag{16}$$

Table IV. Features of the program CARE.

	CARE-1	CARE-2
Application	Epidemiological data (without a natural time ordering)	Animal data (usually with a natural time ordering)
Environment/language	S-plus	C
Data format	Aggregated categorical data	(With covariates) individual capture history (Without covariates) both types of data
Model	Log-linear models Sample coverage approach	Ecological multiplicative and logistic models
Number of samples (t)	$t=2$ to 6	$t \geqslant 2$ for homogeneous models $t \geqslant 5$ is suggested for heterogeneous models
Estimators	All estimators are independent of the ordering of the lists	Some estimators depend on the ordering of the lists

Note that when $t = 3$, (16) reduces to (12) because $tD - \sum\sum A(i,j) = Z_{+11} + Z_{1+1} + Z_{+11}$. The one-step estimator is given by

$$\hat{N}_1 = \frac{D}{\hat{C}} + \frac{1}{t\hat{C}} \sum\sum_{i<j} A(i,j)\hat{\gamma}_{ij} \tag{17}$$

where $\hat{\gamma}_{ij}$ is similarly defined as those in (13a). Standard error estimate and confidence interval can be analogously constructed using a bootstrap method.

5. PROGRAM 'CARE' WITH EXAMPLES

The program CAPTURE, specifically developed for analysing closed ecological data, has not been updated since 1991 [44]. It is felt that an additional program might be needed because some new estimators have been proposed. We have developed a program CARE (for capture-recapture) containing two parts: CARE-1 and CARE-2. The size of the program is about 400 kB. CARE-1 is an S-plus [59] program for analysing epidemiological data; CARE-2, written in C language, calculates various estimates for multiplicative ecological models. The features for these two subprograms are presented in Table IV. The program CARE is available on the first author's website at http://www.stat.nthu.edu.tw/~chao/.

Since our focus here is on epidemiological applications, we only demonstrate in this section the use of CARE-1, but some results from CARE-2 will be given in Section 5.4 for a data with five lists. The reader is referred to the previously-mentioned website for the use of CARE-2. The four data sets mentioned in Section 1 are used for illustration. The HAV data are displayed in Table II (three lists) and the other three sets on spina bifida (three lists), diabetes (four lists) and birth defects (five lists) are shown in Table V.

Table V. Data on spina bifida, diabetes and congenital anomaly.

Spina bifida list			Data
B	D	M	
0	0	0	$Z_{000} = ??$
0	1	0	$Z_{010} = 49$
0	1	1	$Z_{011} = 4$
1	0	0	$Z_{100} = 247$
1	0	1	$Z_{101} = 112$
1	1	0	$Z_{110} = 142$
1	1	1	$Z_{111} = 12$

Diabetes list				Data
1	2	3	4	
0	0	0	0	$Z_{0000} = ??$
0	0	1	0	$Z_{0010} = 182$
0	0	1	1	$Z_{0011} = 8$
0	1	0	0	$Z_{0100} = 74$
0	1	0	1	$Z_{0101} = 7$
0	1	1	0	$Z_{0110} = 20$
0	1	1	1	$Z_{0111} = 14$
1	0	0	0	$Z_{1000} = 709$
1	0	0	1	$Z_{1001} = 12$
1	0	1	0	$Z_{1010} = 650$
1	0	1	1	$Z_{1011} = 46$
1	1	0	0	$Z_{1100} = 104$
1	1	0	1	$Z_{1101} = 18$
1	1	1	0	$Z_{1110} = 157$
1	1	1	1	$Z_{1111} = 58$

Congenital anomaly list					Data
1	2	3	4	5	
0	0	0	0	0	$Z_{00000} = ??$
0	0	0	0	1	$Z_{00001} = 83$
0	0	0	1	0	$Z_{00010} = 97$
0	0	0	1	1	$Z_{00011} = 30$
0	0	1	0	0	$Z_{00100} = 4$
0	0	1	0	1	$Z_{00101} = 3$
0	0	1	1	0	$Z_{00110} = 2$
0	0	1	1	1	$Z_{00111} = 0$
0	1	0	0	0	$Z_{01000} = 37$
0	1	0	0	1	$Z_{01001} = 34$
0	1	0	1	0	$Z_{01010} = 37$
0	1	0	1	1	$Z_{01011} = 23$
0	1	1	0	0	$Z_{01100} = 1$
0	1	1	0	1	$Z_{01101} = 0$
0	1	1	1	0	$Z_{01110} = 3$

Table V. (Continued)

Congenital anomaly list					Data
1	2	3	4	5	
0	1	1	1	1	$Z_{01111} = 0$
1	0	0	0	0	$Z_{10000} = 27$
1	0	0	0	1	$Z_{10001} = 36$
1	0	0	1	0	$Z_{10010} = 22$
1	0	0	1	1	$Z_{10011} = 5$
1	0	1	0	0	$Z_{10100} = 4$
1	0	1	0	1	$Z_{10101} = 5$
1	0	1	1	0	$Z_{10110} = 1$
1	0	1	1	1	$Z_{10111} = 3$
1	1	0	0	0	$Z_{11000} = 19$
1	1	0	0	1	$Z_{11001} = 18$
1	1	0	1	0	$Z_{11010} = 25$
1	1	0	1	1	$Z_{11011} = 8$
1	1	1	0	0	$Z_{11100} = 1$
1	1	1	0	1	$Z_{11101} = 2$
1	1	1	1	0	$Z_{11110} = 5$
1	1	1	1	1	$Z_{11111} = 2$

5.1. *Hepatitis A virus data (three-sample, low sample coverage)*

The analysis procedures for the HAV data given in Table II are the following (the program CARE-1 must be executed in an S-plus environment [59]; what the user needs to input is shown in bold face throughout this section):

1. Insert the CARE disk in a floppy disk drive, say in drive a. Invoke S-plus and type **source**("**a:/care-1.txt**") after a prompt sign, then press the ⟨Enter⟩ key. The following display is shown:

```
CARE-1: for applications to epidemiological data.
This program is used to estimate population size
based on incomplete sources by capture-recapture methods.
The models considered include the log-linear models and the
sample coverage approach. Output includes population size
estimate and its associated standard error as well as a 95%
confidence interval (cil,ciu).

The necessary change in the S-PLUS environment:
* Under the S-PLUS toolbox, please select Options under Main
  Menu -> General Settings -> Computations -> Max Recursion,
  then change the default value 256 to 1024.

Before using this program, please check the following
assumptions:
* Interpretation or definition for the characteristic of the
  target population should be consistent for all data sources.
```

```
      * Closure assumption: the size of the population is
        approximately a constant during the study period.
      * Ascertainable assumption: each case must be ascertainable
        for all sources, although the probability of ascertainment
        is allowed to be heterogeneous.
      * For all sources, identification marks are correctly
        recorded and matched.

Please select:
1: three-source case
2: four-source case
3: five-source case
4: six-source case
5: exit
Selection: 1
```

2. The next step is data entry. CARE-1 can only handle categorical data. Since the HAV data consist of three lists, select 1 (three-source case) as above. Then press the ⟨Enter⟩ key and do the following step-by-step data entry:

```
Your selection is 1 (three-source)
Please key in Z001: 63
Please key in Z010: 55
Please key in Z011: 18
Please key in Z100: 69
Please key in Z101: 17
Please key in Z110: 21
Please key in Z111: 28
```

3. When data entry is finished, press ⟨Enter⟩. After a while, the output is shown as further below. For three-list data, the output includes: (i) estimates based on any pair of samples; this part includes the standard Petersen estimator and the nearly unbiased estimator (the Chapman estimator [10]). Although these two estimates are valid only under the restrictive independence assumption, they are practically useful as a preliminary analysis [22, 47]; (ii) estimates based on various log-linear models; and (iii) estimates obtained from the sample coverage approach. For programming convenience, the lists are consecutively labelled as list 1, 2 and 3. The correspondence to the user's label of lists should be clear.

OUTPUT:

```
Number of identified cases in each list:
       n1        n2        n3
      135       122       126

(1) ESTIMATES BASED ON ANY PAIR OF SAMPLES:

                Petersen      Chapman       se      cil      ciu
     pair(1,2)       336          334       29      289      403
```

pair(1,3)	378	374	36	319	461
pair(2,3)	334	331	30	285	404

Note 1: Refer to Seber (1982, pages 59 and 60) for the Petersen estimator and the Chapman estimator as well as s.e. formula.

Note 2: A log-transformation is used to obtain the confidence interval so that the lower limit is always greater than the number of ascertained. Refer to Chao (1987, Biometrics, 43, 783-791) for the construction of the confidence interval.

(2) ESTIMATES BASED ON LOG-LINEAR MODELS:

	dev.	df	est	se	cil	ciu
independent	24.36	3	388	23	352	442
13/2	24.25	2	393	28	350	461
23/1	21.33	2	413	31	364	488
12/3	21.14	2	416	32	365	494
12/23	13.20	1	527	80	412	735
12/13	19.42	1	464	60	377	622
23/13	19.90	1	452	54	373	592
symmetry	2.05	4	1314	520	685	2899
quasi-sy	0.96	2	1313	520	685	2899
part-qs1	0.03	1	1309	519	682	2891
part-qs2	0.86	1	1306	517	681	2882
part-qs3	0.55	1	1325	528	688	2934
saturated	0.00	0	1313	522	683	2904

DEFINITIONS for the log-linear models:

dev.: deviance statistic for testing goodness of fit.
df: degree of freedom.
est: estimate.
se: asymptotic standard error.
cil: 95% confidence interval lower limit (using a log transformation).
ciu: 95% confidence interval upper limit (using a log transformation).

For the 3-list case, all models are special cases of the following:

$$\log E(Z_{ijk}) = u + u_1 I(i=1) + u_2 I(j=1) + u_3 I(k=1) + u_{12} I(i=j=1) + u_{13} I(i=k=1) + u_{23} I(j=k=1)$$

independent:(independent model) $u_{12} = u_{13} = u_{23} = 0$.
13/2:(model with one interaction) $u_{12} = u_{23} = 0$.
23/1:(model with one interaction) $u_{12} = u_{13} = 0$.
12/3:(model with one interaction) $u_{13} = u_{23} = 0$.
12/23:(model with two interactions) $u_{13} = 0$.
12/13:(model with two interactions) $u_{23} = 0$.
13/23:(model with two interactions) $u_{12} = 0$.

```
symmetry:(symmetry model) u1 = u2 = u3, u12 = u13 = u23.
quasi-sy:(quasi-symmetry model) u12 = u13 = u23.
part-qs1:(partial-quasi-symmetry model) u12 = u23.
part-qs2:(partial-quasi-symmetry model) u12 = u13.
part-qs3:(partial-quasi-symmetry model) u23 = u13.
saturated:(saturated model) no restriction.
```

(3) SAMPLE COVERAGE APPROACH:

	M	D	C^	est	se	cil	ciu
Nhat-0	271	208.667	0.513	407	28	363	472
Nhat	271	208.667	0.513	971	925	369	5290
Nhat-1	271	208.667	0.513	508	40	442	600

parameter estimates:

	u1	u2	u3	r12	r13	r23	r123
Nhat-0	0.33	0.30	0.31	0.21	0.08	0.22	0.73
Nhat	0.14	0.13	0.13	1.89	1.57	1.91	6.35
Nhat-1	0.27	0.24	0.25	0.51	0.34	0.52	1.11

```
      DEFINITIONS for the sample coverage approach:
M: number of individuals ascertained in at least one list.
D: the average of the number of individuals listed in the combination
   of two lists omitting the third.
C^: sample coverage estimate, see (7), or Equation (14) of Chao and Tsay
    (1998).
est: population size estimate.
se: estimated standard error of the population size estimation based on
    1000 bootstrap replications.  Note this s.e. might vary with trials.
cil: 95% confidence interval lower limit (using a log-transformation).
ciu: 95% confidence interval upper limit (using a log-transformation).
Nhat-0: population size estimate for independent samples; see (9), or
        Equation (15) of Chao and Tsay (1998).
Nhat: Population size estimate for sufficiently high sample coverage cases;
      see (12), or Equation (20) of Chao and Tsay (1998).
Nhat-1: One-step population size estimate for low sample coverage cases;
        see (13), or Equation (2.21) of Chao et al. (1996). This estimator is
        suggested for use when the estimated s.e. of Nhat is relatively large.
u1,u2,u3: estimated mean probabilities depending on the estimate of N.
r12,r13,r23,r123: estimated coefficient of covariation (CCV) depending on the
                  estimate of N.
```

For the HAV data, the Petersen and Chapman estimates are in the range of 330 to 380. As discussed in Section 3.2, the Petersen estimator based on two samples is biased downwards (upwards) if these two samples are positively (negatively) dependent. However, the narrow range of these estimates would not indicate the possible direction of dependence at this stage.

The estimated dependence parameters are provided in the output of the sample coverage approach.

The second part of the output includes the results for all possible log-linear models fitted to these data. The outputs show the corresponding deviances and estimates of the total number of infected. The notation and definitions for various models are introduced in Section 4.2 as well as in the output. The independent model produces an estimate of 388, which is close to the results for any two samples. Except for the saturated model, all the log-linear models, which consider local independence only and do not take into account heterogeneity (that is, models PE/Q, QE/P, PQ/E, PQ/QE, PQ/PE and QE/PE), do not fit the data well. All other models, which take heterogeneity only into account (quasi-symmetric and partial quasi-symmetric models) fit well. Those adequate models produce approximately the same estimates of 1300 with an approximate estimated SE of 520. This relatively large estimated SE shows that the data are actually insufficient to model heterogeneous models.

The third part of the output contains the sample coverage approach. The estimators $\hat{N}_0$, $\hat{N}$ and $\hat{N}_1$ derived in equations (9), (12) and (13) correspond to Nhat-0, Nhat and Nhat-1, respectively, in the output. Other statistics can be easily identified in the output because of similar consistent notation. The sample coverage based on (7) is estimated to be

$$\hat{C} = 1 - \frac{1}{3}\left(\frac{Z_{100}}{n_1} + \frac{Z_{010}}{n_2} + \frac{Z_{001}}{n_3}\right) = 1 - \frac{1}{3}\left(\frac{69}{135} + \frac{55}{122} + \frac{63}{126}\right) = 51.3 \text{ per cent}$$

where 69, 55 and 63 are the numbers of singletons. The average of the overlapped cases is equal to $D = 271 - (69 + 55 + 63)/3 = 208.67$. If we ignore the possible dependence between samples, an estimate based on (9) for the HAV data is $\hat{N}_0 = D/\hat{C} = 208.67/0.513 = 407$, which is slightly higher than the estimate of 388 based on the independent log-linear model. The estimator given in (12) is $\hat{N} = 971$, but a large estimated bootstrap SE (925) renders the estimate useless. The estimated SE was calculated by using a bootstrap method based on 1000 replications. We feel these data with a sample coverage estimate of 51 per cent do not contain enough information to correct for undercount. The proposed one-step estimator in equation (13) is $\hat{N}_1 = 508$ with an estimated SE of 40 using 1000 bootstrap replications. The same bootstrap replications produce a 95 per cent confidence interval of (442, 600). We remark that the estimated SE might vary from trial to trial because replications vary in the bootstrap procedures.

It follows from (11) that the CCV measures depend on the value of N. In the output, the CCV estimates based on the three estimates of N show that any two or three samples are positively dependent. As a result, the estimate $\hat{N}_1 = 508$ can only serve as a lower bound. Also, the estimates (Petersen and Chapman's estimates) assuming independence based on two samples should have a negative bias. However, we cannot distinguish which type of dependence (local dependence or heterogeneity) is the main cause of the bias.

In December 1995, the National Quarantine Service of Taiwan conducted a screen serum test for the HAV antibody for all students of the college at which the outbreak of the HAV occurred [1]. After suitable adjustments, they have concluded that the final figure of the number infected was about 545. Thus this example presents a very valuable data set with the advantage of a known true parameter. Our estimator $\hat{N}_1$ does provide a satisfactory lower bound. This example shows the need for undercount correction and also the usefulness of the capture-recapture method in estimating the number of missing cases.

5.2. Spina bifida data (three-sample)

The data set on spina bifida [2, 3] was reproduced in Table V. Three lists were collected: birth certificates (B-list); death certificates (D-list), and medical rehabilitation files (M-list). There were 513, 207 and 188 cases, respectively, on B-, D- and M-lists and in total 626 ascertained cases. After the data entry, part of the output shows:

```
OUTPUT:
Number of identified cases in each list:
       n1        n2        n3
      513       207       188
(1) ESTIMATES BASED ON ANY PAIR OF SAMPLES:
                    Petersen      Chapman       se      cil      ciu
      pair(1,2)          690          689       23      651      743
      pair(1,3)          778          776       35      718      857
      pair(2,3)         2432         2311      498     1554     3556
(2) ESTIMATES BASED ON LOG-LINEAR MODELS:
                        dev.     df      est      se      cil      ciu
      independent      58.35      3      764      21      728      812
            13/2       58.09      2      756      25      715      816
            23/1        3.86      2      731      17      702      771
            12/3       46.85      2      831      37      770      919
           12/23        0.00      1      758      26      716      820
           12/13       37.50      1     1361     396      899     2602
           23/13        0.67      1      711      18      683      754
        symmetry      370.66      4      658      13      640      696
        quasi-sy       29.01      2      649      10      636      678
        part-qs1       29.00      1      649      10      636      679
        part-qs2        3.67      1      762      86      670     1051
        part-qs3       15.79      1      659      14      641      700
       saturated        0.00      0      763      87      670     1053

(3) SAMPLE COVERAGE APPROACH:
                   M         D      C^       est      se      cil      ciu
      Nhat-0     626   507.333   0.654       775      22      738      826
        Nhat     626   507.333   0.654       752      36      699      844
      Nhat-1     626   507.333   0.654       767      33      716      848

  parameter estimates:
                   u1       u2       u3       r12       r13       r23      r123
      Nhat-0     0.66     0.27     0.24      0.12      0.00     -0.68     -0.08
        Nhat     0.68     0.28     0.25      0.09     -0.03     -0.69     -0.03
      Nhat-1     0.67     0.27     0.25      0.11     -0.01     -0.68     -0.06
```

Note that the Petersen and Chapman estimates based on the D- and M-lists are substantially higher than the other two estimates. This implies that a possible negative dependence exists between these two lists. Using three samples, Regal and Hook [3] showed that there is a strong

negative dependence between the D- and M-lists, moderate positive dependence between the B- and D-lists and weak dependence for the B- and M-lists. Regal and Hook indicated that an adequate model is DM/BD, which gives a population size estimate of 758 with a 95 per cent confidence interval of (707, 809), but they also commented that this confidence interval might be artificially narrow. A pairwise DM/BD/BM model yields an estimate of 763 with a 95 per cent confidence interval of (590, 936). Their confidence bounds are slightly different from those in the output because a log-transformation is adopted in our approach.

The sample coverage estimate for this data set is 65.4 per cent, and $D = 507.33$, so an estimate without taking into account the possible dependence is $\hat{N}_0 = 775$. The proposed estimate $\hat{N} = 752$ (SE = 36) and the corresponding 95 per cent confidence interval is (699, 844) using 1000 bootstrap replications. These data are sufficient to produce a reliable population size estimate. Both the log-linear and our approaches produce very close estimates. Our interval is slightly wider than the one obtained by the DM/BD model but much narrower than that obtained by the pairwise DM/BD/BM model. Based on (11) and an estimated population size of 752, the output shows that $\hat{\gamma}_{DM} = -0.69$, $\hat{\gamma}_{BD} = 0.09$ and $\hat{\gamma}_{BM} = -0.03$. These estimated CCV values support the finding of Regal and Hook about the dependencies between two samples. Here the CCV estimates represent the mixed effects of two types of dependencies.

5.3. *Diabetes data (four-sample)*

The data given in Table V on diabetes were collected and discussed by Bruno *et al.* [4] and IWGDMF [8, 9]. The purpose of collecting these data was to estimate the number of diabetes patients in a community in Italy based on the following four lists: diabetic clinic and/or family physician visits (list 1, 1754 cases); hospital discharges (list 2, 452 cases); prescriptions (list 3, 1135 cases), and purchases of regent strips and insulin syringes (list 4, 173 cases). A total of 2069 cases were identified. When the number of samples is at least four, there are many available log-linear models. We provide six selections in the program. The basic models include all the estimates based on any pair of samples, some selected log-linear models and the sample coverage approach. The other five selections include different types of log-linear models. This program does not provide a model selection procedure and the users have to select their own model. In the following, we show the procedure for analysing the diabetes data using basic models as well as models which include five and six two-factor interactions with/without heterogeneity:

```
Please select:
1: three-source case
2: four-source case
3: five-source case
4: six-source case
5: exit
Selection: 2
your selection is  2 (four-source)
Please key in Z0001: 10
Please key in Z0010: 182
Please key in Z0011: 8
Please key in Z0100: 74
Please key in Z0101: 7
```

```
        Please key in Z0110: 20
        Please key in Z0111: 14
        Please key in Z1000: 709
        Please key in Z1001: 12
        Please key in Z1010: 650
        Please key in Z1011: 46
        Please key in Z1100: 104
        Please key in Z1101: 18
        Please key in Z1110: 157
        Please key in Z1111: 58
Please select:
   1: basic models (including pair-sample models, an independent model,
      symmetric model, quasi-symmetric model, models with one 3-factor
      interaction and the sample coverage estimates)
   2: models with one 2-factor interaction w/o H1, H2
      (H1: the first order heterogeneity, i.e., all 2-factor interactions
       are identical)
      (H2: the second order heterogeneity, i.e., all 3-factor interactions
       are identical)
   3: models with two 2-factor interactions w/o H1, H2
   4: models with three 2-factor interactions w/o H1, H2
   5: models with four 2-factor interactions w/o H1, H2
   6: models with five & six 2-factor interactions w/o H1, H2
   7: exit
Selection: 1

OUTPUT:

Number of identified cases in each list:
         n1        n2        n3        n4
       1754       452      1135       173

(1) ESTIMATES BAS
ED ON ANY PAIR OF SAMPLES:
                 Petersen    Chapman    se   cil   ciu
      pair(1,2)      2353       2351    58  2250  2478
      pair(1,3)      2185       2185    22  2146  2233
      pair(1,4)      2264       2261    88  2117  2468
      pair(2,3)      2060       2057    77  1922  2224
      pair(2,4)       806        803    47   725   913
      pair(3,4)      1558       1555    67  1445  1712

(2) ESTIMATES BASED ON LOG-LINEAR MODELS:
                       dev.      df      est      se      cil     ciu
      independent    217.48      10     2251      19     2217    2292
            123/4    165.76       6     2185      40     2130    2291
```

124/3	92.23	6	2247	21	2211	2293
134/2	154.38	6	2386	45	2309	2487
234/1	55.24	6	2283	22	2244	2331
H1	105.63	9	2669	83	2528	2854
symmetry	3156.50	11	2197	50	2130	2338
quasi-sy	93.95	8	2239	68	2148	2431
saturated	0.00	0	5367	2771	2856	15883

(3) SAMPLE COVERAGE APPROACH:

	M	D	C^	est	se	cil	ciu
Nhat-0	2069	1825.25	0.803	2272	26	2226	2330
Nhat	2069	1825.25	0.803	2609	81	2472	2792
Nhat-1	2069	1825.25	0.803	2458	50	2372	2568

parameter estimates:

	u1	u2	u3	r12	r13	r14	r23	r24	r34
Nhat-0	0.77	0.20	0.50	0.08	-0.03	0.04	0.00	0.10	1.82
Nhat	0.67	0.17	0.44	0.07	0.11	0.19	0.15	0.27	2.24
Nhat-1	0.71	0.18	0.46	0.07	0.04	0.12	0.09	0.19	2.05

```
   If you want to continue to fit other models
   please select:
 1: basic models (including pair-sample models, an independent model,
     symmetric model, quasi-symmetric model, models with one 3-factor
     interaction and the sample coverage estimates)
 2: models with one 2-factor interaction w/o H1, H2
     (H1: the first order heterogeneity, i.e., all 2-factor interactions
      are identical)
     (H2: the second order heterogeneity, i.e., all 3-factor interactions
      are identical)
 3: models with two 2-factor interactions w/o H1, H2
 4: models with three 2-factor interactions w/o H1, H2
 5: models with four 2-factor interactions w/o H1, H2
 6: models with five & six 2-factor interactions w/o H1, H2
 7: exit
```

Selection: **6**

OUTPUT:

Estimates based on log-linear models:

	dev.	df	est	se	cil	ciu
12/13/14/23/24	52.03	5	2648	122	2455	2939
12/13/14/23/34	135.33	5	2630	119	2440	2916
12/13/14/24/34	16.74	5	2637	116	2452	2912
12/13/23/24/34	7.62	5	2771	146	2538	3120

	12/14/23/24/34	53.10	5	2271	23	2230	2323
	13/14/23/24/34	13.72	5	2562	75	2435	2733
H1	12/13/14/23/24	7.05	4	2790	153	2547	3155
H1	12/13/14/23/34	7.05	4	2790	153	2547	3155
H1	12/13/14/24/34	7.05	4	2790	153	2547	3155
H1	12/13/23/24/34	7.05	4	2790	153	2547	3155
H1	12/14/23/24/34	7.05	4	2790	153	2547	3155
H1	13/14/23/24/34	7.05	4	2790	153	2547	3155
H1 H2	12/13/14/23/24	0.92	3	4501	1319	2968	8647
H1 H2	12/13/14/23/34	0.92	3	4501	1319	2968	8647
H1 H2	12/13/14/24/34	0.92	3	4501	1319	2968	8647
H1 H2	12/13/23/24/34	0.92	3	4501	1319	2968	8647
H1 H2	12/14/23/24/34	0.92	3	4501	1319	2968	8647
H1 H2	13/14/23/24/34	0.92	3	4501	1319	2968	8647
	12/13/14/23/24/34	7.05	4	2790	153	2547	3155

Bruno *et al.* [4] found that the log-linear model 12/13/23/24/34 fits the data well and presented an estimate of 2771 with a 95 per cent confidence interval of (2492, 3051). Their estimate is shown in the above output. The use of a log-transformation shifts their interval rightward and results in little increase in interval length. When they further stratified for the pattern of treatment (dietary control, hypoglycaemia agents and insulin), an estimate of 2586 was obtained with a 95 per cent interval of (2341, 2830). The two review papers by IWGDMF [8, 9] analysed the data by including heterogeneity terms to several proper log-linear models and selected the final model by the Akaike information criterion. They obtained an estimate of 2834 based on the stratified data.

The sample coverage for these data is estimated to be 80.3 per cent. Since the coverage estimate is sufficiently high, a precise estimate is expected. The proposed estimate is $\hat{N} = 2609$ with an estimated SE of 81 using 1000 bootstrap replications. Positive dependencies exist in any two samples as shown in the output. The corresponding 95 per cent confidence interval for $\hat{N}$ is (2472, 2792). If the data are analysed within each stratum, we have the sum of the three estimates as $\hat{N} = 2559$, which is very close to the unstratified result.

5.4. *Infants' congenital anomaly data (five-sample)*

For this data set, Wittes *et al.* [5] obtained an estimate of 638 (SE 15) for the total number of cases under an independent assumption. This result can be seen from the second part of our output (see below) for the log-linear model approach. Fienberg [6] further modelled the possible dependencies between samples and fitted a log-linear model (12/14/25/34) to the data. He found that all the four interactions were highly significant and obtained an estimate of 634 (SE 18). Fienberg's estimate is very close to the result under independence because there are both positive and negative dependence-effects in the model; see the CCV estimates below. Part of the output from the program CARE-1 after data entry is shown below.

```
Number of identified in each list:
      n1      n2      n3      n4      n5
     183     215      36     263     252
(1) ESTIMATES BASED ON ANY PAIR OF SAMPLES:
```

	Petersen	Chapman	se	cil	ciu
pair(1,2)	492	490	32	437	565
pair(1,3)	286	283	31	240	368
pair(1,4)	678	674	53	587	796
pair(1,5)	584	581	40	515	674
pair(2,3)	553	532	99	392	797
pair(2,4)	549	547	30	498	617
pair(2,5)	623	620	41	552	714
pair(3,4)	592	574	96	437	830
pair(3,5)	605	584	103	438	860
pair(4,5)	933	927	78	796	1106

(2) ESTIMATES BASED ON LOG-LINEAR MODELS:

	dev.	df	est	se	cil	ciu
independent	93.45	25	638	15	613	673
symmetry	396.96	26	756	424	556	3064
quasi-sy	87.69	22	806	521	560	3647

(3) SAMPLE COVERAGE APPROACH:

	M	D	C^	est	se	cil	ciu
Nhat0	537	487.4	0.774	630	15	604	664
Nhat	537	487.4	0.774	659	35	607	750
Nhat-1	537	487.4	0.774	649	27	608	714

parameter estimates:

	u1	u2	u3	u4	u5
Nhat-0	0.29	0.34	0.06	0.42	0.40
Nhat	0.28	0.33	0.05	0.40	0.38
Nhat-1	0.28	0.33	0.06	0.41	0.39

	r12	r13	r14	r15	r23	r24	r25	r34	r35
Nhat-0	0.28	1.20	-0.07	0.08	0.14	0.15	0.01	0.06	0.04
Nhat	0.34	1.30	-0.03	0.13	0.19	0.20	0.06	0.11	0.09
Nhat-1	0.32	1.27	-0.04	0.11	0.17	0.18	0.04	0.10	0.07

	r45
Nhat-0	-0.33
Nhat	-0.29
Nhat-1	-0.30

First notice that two of the ten pairwise estimates lie below the other values, especially the estimate using list 1 and list 3. This shows strong evidence of positive dependence between these two lists. Also, negative dependence arises between samples 4 and 5 as the estimate using this pair of samples is much higher. The other estimates are in the range of 550 to 680.

For the sample coverage approach, the overlap fraction is estimated by the sample coverage estimate, which is 77.4 per cent. Our proposed estimator based on equation (16) is 659 (SE

35) with a 95 per cent confidence interval of (607, 750). Our estimator that incorporates both types of dependencies agrees well with the previous findings, but the variation is larger due to estimating more dependence parameters. The CCV estimates show that there are relatively large positive values (γ_{12} and γ_{13}) and negative value (γ_{45}). The dependence between samples 1 and 3 is significantly higher than the others, but it was not included in Fienberg's model.

For this five-sample data set, we can illustrate the use of heterogeneous ecological models, although those models are typically applied to situations with identical trapping methods. As discussed in Section 4.1, only the following models are potentially useful: multiplicative model $\mathbf{M}_{th}$; logistic model $\mathbf{M}_{th}$ (that is, the Rasch model); model $\mathbf{M}_h$, and model $\mathbf{M}_t$ (for the latter two models, the multiplicative and logistic types of models are equivalent). Model $\mathbf{M}_t$ is equivalent to the independent log-linear model and this is the model considered by Wittes *et al.* [5]. For the Rasch model, it is equivalent to a quasi-symmetric model with some constraints. From the second part of the output, the estimate under a quasi-symmetric model is 806 (SE 521). However, the estimated SE is large so that the model is unlikely to be useful. For model $\mathbf{M}_h$, the first-order and second-order jackknife estimators (Burnham and Overton [46]) are, respectively, 735 (SE 19) and 800 (SE 27). The interpolated jackknife combining the first- and second-order jackknife is 772 (SE 49). The two estimators proposed by Lee and Chao [43] are 770 (SE 32) and 641 (SE 21) under model $\mathbf{M}_h$, and 789 (SE 36) and 654 (SE 25) under model $\mathbf{M}_{th}$. Although these two heterogeneous models do not consider the possible local dependence, it is interesting to notice that one of the estimates under each model is close to the proposed estimate that considers two types of dependencies. All the estimates and SEs for models $\mathbf{M}_{th}$ and $\mathbf{M}_h$ were obtained using our program CARE-2.

6. REMARKS AND DISCUSSION

Three classes of capture-recapture models have been reviewed in this tutorial: ecological models; log-linear models, and the sample coverage approach. Most ecological models allowing for heterogeneous capture probabilities are recommended only when there are at least five trapping samples, whereas the other two approaches are mainly useful for two to five samples. We have focused on the latter two models for epidemiological applications and demonstrated the use of the program CARE developed by the authors.

Hook and Regal [60, 61] presented 17 recommendations for the use of the capture-recapture method in epidemiology. There are several basic assumptions that should be fulfilled or checked to validate the application of the method. In addition to the closure assumption, a basic assumption is an explicit definition or interpretation of the 'target population'. Gutteridge and Collin, in a prevalence study of physical disability [62], reported that two sources might have different interpretations of disability and its severity. Thus the 'target population' for two sources might become inconsistent. Another basic assumption is that all identification 'marks' should be correctly recorded and matched. Although in most epidemiological studies this assumption obviously can be fulfilled, in reality it might be an impediment in developing countries, as indicated by Black and McLarty [63].

An implicit assumption is that the joint 'capture' probability for any individual in *all* lists should be positive so that overlap information can be obtained. This implies that any individual must have a positive probability to be ascertained by any source and unascertainment is purely a 'random zero' (missing due to small chance), not a 'structural zero' (missing due to

impossibility). If some cases are systematically missed by one or more sources, then there is no overlap information and those 'uncatchable' individuals cannot be included in our target population and should be treated separately. An extreme example of this situation can be seen where the first list collected cases from a certain area, whereas the other list collected cases from another disjoint area. There is no way to get source intersection, consequently capture-recapture cannot be used to estimate the total number of cases in the combined whole area. The estimating target is actually the size of those jointly 'ascertainable' or 'catchable' individuals by *all* sources. Therefore, two complementary lists collected from disjoint areas cannot be utilized as two separate lists and they should be combined into a joint list.

We also note that another limitation of the capture-recapture methods is that sufficiently high overlapping information is required to produce reliable population size estimates and to model dependence among samples. Otherwise, Coull and Agresti [40] demonstrated that the likelihood functions under some random-effect models for sparse information might become flat and the resulting estimates based on equivalent log-linear models are likely to become unstable. Chao *et al.* [29] also showed in a sample coverage approach that a large variation might be associated with the resulting estimator due to insufficient overlap. In such cases, we have proposed a plausible lower bound given in equations (13) and (17) for positively dependent samples. We feel that a precise bound is of more practical use than an imprecise point estimate.

As indicated by the IWGDMF [8, 9], the log-linear model approach has many advantages as follows: (a) all models are under a unified framework; (b) model selection can be easily implemented and carried out in a flexible fashion, based on data and guided by prior information; (c) tests are available for comparing models; (d) dependence can be incorporated by adding proper interactions; and (e) all inference is within the mainstream of statistical data analysis. An untestable assumption involved is that the highest-order interaction does not exist. The IWGDMF commented that the existence of heterogeneity in three-list data might result in the lack of a reliable estimate. Another concern is that two equally fitted models might produce quite different estimates [reference 40, p. 299]. As the number of lists increases, the number of adequate models increases rapidly and thus model selection causes further problems [23].

The sample coverage approach provides an alternative approach, which makes use of overlap information to incorporate source-dependence in the estimation. The advantages for this approach include the following: (a) overlap information can be quantified by the estimated sample coverage; (b) dependence among samples can be quantified by estimated CCVs, and thus can be detected by data; (c) no model selection or model comparison is needed; and (d) no further difficulty arises when the number of lists increases. The program CARE-1 can deal with data up to six lists and can be easily extended to handle data with more than six lists. Nevertheless, there is an untestable and quite complicated assumption. Recall that this approach is mainly derived from an expansion (equation (10)) and its generalization; the assumption is that the limiting value of the remainder term of that expansion tends to zero. This assumption is satisfied under the gamma type of heterogeneity, but the robustness of the resulting estimator to the departure from the gamma distribution needs further investigation [29, 53].

Some major problems with the use of capture-recapture models have been raised by previous authors [14, 25–28] and have been encountered in our consulting services with researchers in health science. Their problems (listed below) represent the major concerns among epidemiologists about the method. To clarify the use of this methodology and to enhance the

understanding of its application, we focus on the following principal concerns and include some discussion from a statistical point of view.

6.1. Is it necessary to have at least a random sample?

If a random sample could be obtained, then two samples would suffice for estimating population size. A random sample implies that all individuals in the population have the same probability to be ascertained in that list. As we discussed in Section 4.3, the usual Petersen estimator is valid if the second sample is random. No correlation bias arises even when the first non-random sample is highly selective or extremely heterogeneous. This can be understood intuitively. For example, if we only tag large fish in the first sample, but fish of any size are captured in an equally likely manner in the recapture sample, then local dependence clearly does not arise because marked and unmarked have identical chances of being caught. The marked rate in the sample is approximately the rate in the population, which justifies the use of the Petersen estimator (see Section 3.2). Moreover, when only heterogeneity is present, correlation bias vanishes if either sample is random. A random tagging in the first sample means that the tag rates are nearly the same in any groups of different sizes. Even if we only catch large fish in the recapture sample, the fraction of tagged individuals in the sample is still approximately equal to the fraction tagged in the whole population. These findings could also be theoretically justified by the definition of CCV in the sample coverage approach (Section 4.3). One of the motivations for developing models incorporating dependence is that a random sample is almost unfeasible for animal studies. Animals cannot be drawn in a randomized manner. An advantage of using more sophisticated dependent models to estimate population size is that a random sample is not necessary.

6.2. Is it better to use an identical ascertainment method for all lists?

As we have discussed in Section 2.2, identical trapping methods are usually used in animal studies. The advantages from a statistical point of view are threefold: (a) dependence patterns between samples are similar, so one or two parameters are sufficient to model dependence no matter how many samples are taken, so simpler models can be adopted for modelling heterogeneous populations; (b) as more samples are conducted, overlap information is generally increased without inducing additional dependence measures; and (c) finally, systematic missing patterns are unlikely to occur and all animals are 'jointly catchable'. However, there are disadvantages: (a) possible local dependence due to a behavioural response to identical trapping experience might be induced; and (b) dependence due to heterogeneity arises because of strong correlation between two sets of similar capture probabilities. To reduce correlation bias, different trapping methods have been proposed by researchers, especially in fishery science [reference 10, p. 86]. Thus, identical ascertainment methods are not necessary, but we need to model more dependence measures and must be cautious about the possible missing patterns or structural zeros if different surveys are applied.

6.3. The traditional assumption of independence

As we have explained for the two-list cases in Section 3.2, this is indeed a problem unless one is willing to assume some value for the two-sample interaction. When there are three or more sources, the log-linear models and the sample coverage approach provide viable

ways to model dependence. Previous studies [8, 9, 29, 53] have shown in many cases that the performance of the two approaches is encouraging.

6.4. Almost zero probability of being captured by any source

As we have remarked before, those 'uncatchable' cases cannot be included in our target 'population' and should be treated separately. We can only estimate the size of a subpopulation that contains only catchable individuals.

6.5. Heterogeneity among individuals ('variable catchability')

This problem has been discussed extensively in animal population studies. Stratified analysis has been suggested in the literature. For example, data for males and females are treated separately, or the data can be stratified by the type of treatment, such as in the case of data on diabetes. However, even in a stratum, residual heterogeneity may still exist. As indicated before, heterogeneity among individuals induces possible source dependence. Therefore, the heterogeneity problem can be partially solved by proper adjustment for dependence. We emphasize that heterogeneity does not always induce dependence and two heterogeneous samples may be independent (see Section 4.3).

6.6. How many sources are needed?

Hay [64], Chang *et al.* [65] and Ismail *et al.* [66] had some interesting observations and suggestions regarding the number of sources used in capture-recapture studies. When identical trapping methods are used as in ecological studies, more individuals would be caught, and overlap information generally would increase without inducing more dependence measures. As a result, a more precise estimator may be produced. In the log-linear model approach, the higher order interaction is likely to be less significant. Thus the basic assumption of there being no highest order interaction is more reasonable when there are more lists. However, in health science, different ascertainments are used and thus more dependence parameters are involved as the number of lists is increased. Moreover, the probability of producing structural zero is increased as there are more cells in the data. Increasing the number of lists often costs more and requires additional effort, but it does not necessarily yield better results, especially when different ascertainment methods are applied. We would thus recommend that only three or four samples be used unless similar types of identification surveys are conducted.

In summary, capture-recapture models provide a potentially useful method to estimate population size in epidemiological studies but there are assumptions and limitations to this approach. The four data sets discussed in this paper have provided examples to show the usefulness of the capture-recapture analysis in assessing the extent of incomplete ascertainment. Efforts are needed to study the relative merits of the existing models and to provide practical guidelines. More collaboration is certainly needed between epidemiologists and statisticians to pursue additional methodological and conceptual research work.

ACKNOWLEDGEMENTS

This paper is an expanded revision of our earlier manuscript on analysing the hepatitis A virus data in Taiwan. We thank all reviewers for providing helpful comments and suggestions on the earlier versions and for pointing out some recent publications (references [28, 47, 60] and [61]). We also thank one

reviewer and the editors (Professors Machin and D'Agostino) for having suggested expansion and resubmission as a tutorial paper. This research was supported by the National Science Council of Taiwan.

REFERENCES

1. Chao DY, Shau WY, Lu CWK, Chen KT, Chu CL, Shu HM, Horng CB. A large outbreak of hepatitis A in a college school in Taiwan: associated with contaminated food and water dissemination. Epidemiology Bulletin, Department of Health, Executive Yuan, Taiwan Government, 1997.
2. Hook EB, Albright SG, Cross PK. Use of Bernoulli census and log-linear methods for estimating the prevalence of spina bifida in livebirths and the completeness of vital record reports in New York State. *American Journal of Epidemiology* 1980; **112**:750–758.
3. Regal RR, Hook EB. The effects of model selection on confidence intervals for the size of a closed population. *Statistics in Medicine* 1991; **10**:717–721.
4. Bruno GB, Biggeri A, LaPorte RE, McCarty D, Merletti F, Pagono G. Application of capture-recapture to count diabetes. *Diabetes Care* 1994; **17**:548–556.
5. Wittes JT, Colton T, Sidel VW. Capture-recapture methods for assessing the completeness of cases ascertainment when using multiple information sources. *Journal of Chronic Diseases* 1974; **27**:25–36.
6. Fienberg SE. The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* 1972; **59**:591–603.
7. Hook EB, Regal RR. Accuracy of alternative approaches to capture-recapture estimates of disease frequency: internal validity analysis of data from five sources. *American Journal of Epidemiology* 2000; **152**:771–779.
8. International Working Group for Disease Monitoring and Forecasting (IWGDMF). Capture-recapture and multiple-record systems estimation I: history and theoretical development. *American Journal of Epidemiology* 1995; **142**:1047–1058.
9. International Working Group for Disease Monitoring and Forecasting (IWGDMF). Capture-recapture and multiple-record systems estimation II: application in human diseases. *American Journal of Epidemiology* 1995; **142**:1059–1068.
10. Seber GAF. *The Estimation of Animal Abundance*, 2nd edn. Griffin: London, 1982.
11. Seber GAF. A review of estimating animal abundance. *Biometrics* 1986; **42**:267–292.
12. Seber GAF. A review of estimating animal abundance II. *International Statistical Review* 1992; **60**:129–166.
13. Schwarz CJ, Seber GAF. A review of estimating animal abundance III. *Statistical Science* 1999; **14**:427–456.
14. Schouten LJ, Straatman H, Kiemeney LALM, Gimbrere CHF, Verbeek ALM. The capture-recapture method for estimation of cancer registry completeness; a useful tool? *International Journal of Epidemiology* 1994; **23**:1111–1116.
15. Hook EB, Regal RR. Validity of Bernoulli census, log-linear, and truncated binomial models for correction for underestimates in prevalence studies. *American Journal of Epidemiology* 1982; **116**:168–176.
16. LaPorte RE, McCarty, DJ, Tull ES, Tajima N. Counting birds, bees and NCDs. *Lancet* 1992; **339**:494.
17. Darroch JN. The Multiple-Recapture Census I. Estimation of a closed population. *Biometrika* 1958; **45**: 343–359.
18. Sekar C, Deming WE. On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association* 1949; **44**:101–115.
19. Wittes JT, Sidel VW. A generalization of the simple capture-recapture model with applications to epidemiological research. *Journal of Chronic Diseases* 1968; **21**:287–301.
20. Wittes JT. Applications of a multinomial capture-recapture method to epidemiological data. *Journal of the American Statistical Association* 1974; **69**:93–97.
21. McCarty DJ, Tull ES, Moy CS, Kwoh CK, LaPorte RE. Ascertained corrected rates: Applications of capture-recapture methods. *International Journal of Epidemiology* 1993; **22**:559–565.
22. Hook EB, Regal RR. The value of capture-recapture methods even for apparently exhaustive surveys: the need for adjustment for source of ascertainment intersection in attempted complete prevalence studies. *American Journal of Epidemiology* 1992; **135**:1060–1067.
23. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitation. *Epidemiological Reviews* 1995; **17**:243–264.
24. Chao A. Capture-recapture. In *Encyclopedia of Biostatistics*, Armitage P, Colton T (eds). Wiley: New York, 1998.
25. Kiemeney LALM, Schouten LJ, Straatman H. Ascertainment corrected rates (Letter to Editor). *International Journal of Epidemiology* 1994; **23**:203–204.
26. Desenclos JC, Hubert B. Limitations to the universal use of capture-recapture methods. *International Journal of Epidemiology* 1994; **23**:1322–1323.

27. Papoz L, Balkau B, Lellouch J. Case counting in epidemiology: limitation of methods based on multiple data sources. *International Journal of Epidemiology* 1996; **25**: 474–477.
28. Cormack RM. Problems with using capture-recapture in epidemiology: an example of a measles epidemic. *Journal of Clinical Epidemiology* 1999; **52**:909–914.
29. Chao A, Tsay PK, Shau WY, Chao DY. Population size estimation for capture-recapture models with applications to epidemiological data. *Proceedings of Biometrics Section, American Statistical Association* 1996; 108–117.
30. Lazarsfeld PF, Henry NW. *Latent Structure Analysis*. Houghton Mifflin: Boston, 1968.
31. Hook EB, Regal RR. Effects of variation in probability of ascertainment by sources ('variable catchability') upon 'capture-recapture' estimates of prevalence. *American Journal of Epidemiology* 1993; **137**:1148–1166.
32. Darroch JN, Fienberg SE, Glonek GFV, Junker BW. A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association* 1993; **88**:1137–1148.
33. Pollock KH. Modeling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife population: past, present, and future. *Journal of the American Statistical Association* 1991; **86**:225–238.
34. Otis DL, Burnham KP, White GC, Anderson DR. Statistical inference from capture data on closed animal populations. *Wildlife Monographs* 1978; **62**:1–135.
35. White GC, Anderson DR, Burnham KP, Otis DL. *Capture-Recapture and Removal Methods for Sampling Closed Populations*. Los Alamos National Lab, LA-8787-NERP: Los Alamos, New Mexico, USA, 1982.
36. Huggins RM. On the statistical analysis of capture experiments. *Biometrika* 1989; **76**:133–140.
37. Alho JM. Logistic regression in capture-recapture models. *Biometrics* 1990; **46**:623–635.
38. Huggins RM. Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics* 1991; **47**:725–732.
39. Rasch G. On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Neyman J (ed.). University of California Press: 1961; 321–333.
40. Coull BA, Agresti A. The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics* 1999; **55**:294–301.
41. Lloyd CJ, Yip P. A unification of inference for capture-recapture studies through martingale functions. In *Estimating Equations*, Godambe VP (ed.). Clarendon Press: Oxford, 1991; 65–88.
42. Pledger S. Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics* 2000; **56**:434–442.
43. Lee SM, Chao A. Estimating population size via sample coverage for closed capture-recapture models. *Biometrics* 1994; **50**:88–97.
44. Rexstad E, Burnham KP. *User's Guide for Interactive Program CAPTURE*. Colorado Cooperative Fish and Wildlife Research Unit: Fort Collins, 1991.
45. Bishop YMM, Fienberg SE, Holland PW. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press: Cambridge, Mass., 1975.
46. Burnham KP, Overton WS. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* 1978; **65**: 625–633.
47. Fienberg SE, Johnson MS, Junker BW. Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of Royal Statistical Society, Series A* 1999; **162**:383–405.
48. Chao A, Lee SM, Jeng SL. Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics* 1992; **48**:201–216.
49. Cormack RM. Loglinear models for capture-recapture. *Biometrics* 1989; **45**:395–413.
50. Agresti A. Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics* 1994; **50**:494–500.
51. Lloyd CJ. *Statistical Analysis of Categorical Data*. Wiley: New York, 1999.
52. Norris JL, Pollock KH. Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics* 1996; **52**:639–649.
53. Chao A, Tsay PK. A sample coverage approach to multiple-system estimation with application to census undercount. *Journal of the American Statistical Association* 1998; **93**: 283–293.
54. Tsay PK, Chao A. Population size estimation for capture-recapture models with applications to epidemiological data. *Journal of Applied Statistics* 2001; **28**:25–36.
55. Good IJ. The population frequencies of species and the estimation of population parameters. *Biometrika* 1953; **40**:237–264.
56. Bunge J, Fitzpatrick M. Estimating the number of species: recent developments. *Journal of the American Statistical Association* 1993; **88**:364–373.
57. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman and Hall: New York, 1993.
58. Chao A. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 1987; **43**:783–791.

59. MathSoft. *S-PLUS User's Manual, Version 4.0*. MathSoft, Inc.: Seattle, WA, 1997.
60. Hook EB, Regal RR. Recommendations for presentation and evaluation of capture-recapture estimates in epidemiology. *Journal of Clinical Epidemiology* 1999; **52**:917–926.
61. Hook EB, Regal RR. On the need for a 16th and 17th recommendation for capture-recapture analysis. *Journal of Clinical Epidemiology* 2000; **53**:1275–1277.
62. Gutteridge W, Collin C. Capture-recapture technique: quick and cheap (Letter). *British Medical Journal* 1994; **308**:531.
63. Black JFP, McLarty DG. Capture-recapture technique: difficult to use in developing countries (Letter). *British Medical Journal* 1994; **308**:531.
64. Hay G. The selection from multiple data sources in epidemiological capture-recapture studies. *Statistician* 1997; **46**:515–520.
65. Chang YF, LaPorte RE, Aaron DJ, Songer TJ. The importance of source selection and pilot study in the capture-recapture application. *Journal of Clinical Epidemiology* 1999; **52**:927–928.
66. Ismail AA, Beeching NJ, Gill GV, Bellis MA. How many data sources are needed to determine diabetes prevalence by capture-recapture? *International Journal of Epidemiology* 2000; **29**:536–541.

1.2 Adjustment Methods

TUTORIAL IN BIOSTATISTICS

PROPENSITY SCORE METHODS FOR BIAS REDUCTION IN THE COMPARISON OF A TREATMENT TO A NON-RANDOMIZED CONTROL GROUP

RALPH B. D'AGOSTINO, Jr.*

Department of Public Health Sciences, Section on Biostatistics, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157-1063, U.S.A.

SUMMARY

In observational studies, investigators have no control over the treatment assignment. The treated and non-treated (that is, control) groups may have large differences on their observed covariates, and these differences can lead to biased estimates of treatment effects. Even traditional covariance analysis adjustments may be inadequate to eliminate this bias. The propensity score, defined as the conditional probability of being treated given the covariates, can be used to balance the covariates in the two groups, and therefore reduce this bias. In order to estimate the propensity score, one must model the distribution of the treatment indicator variable given the observed covariates. Once estimated the propensity score can be used to reduce bias through matching, stratification (subclassification), regression adjustment, or some combination of all three. In this tutorial we discuss the uses of propensity score methods for bias reduction, give references to the literature and illustrate the uses through applied examples. © 1998 John Wiley & Sons, Ltd.

INTRODUCTION

Observational studies occur frequently in medical research. In these studies, investigators have no control over the treatment assignment. Therefore, large differences on observed covariates in the two groups may exist, and these differences could lead to biased estimates of treatment effects. The propensity score for an individual, defined as the conditional probability of being treated given the individual's covariates, can be used to balance the covariates in the two groups, and thus reduce this bias. The propensity score has been used to reduce bias in observational studies in many fields. In particular, there are good recent examples in the literature where propensity scores were discussed in either applied statistical journals[1-7] or in medical journals.[8-21] Topics discussed in these articles come from a variety of fields including epidemiology, health services research, economics and social sciences.

* Correspondence to: Ralph B. D'Agostino, Jr, Department of Public Health Sciences, Section on Biostatistics, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157-1063, U.S.A. E-mail: rdagosti@rc.phs.bgsm.edu

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
 ISBN: 0-470-02365-1

In a randomized experiment, the randomization of units (that is, subjects) to different treatments guarantees that on average there should be no systematic differences in observed or unobserved covariates (that is, bias) between units assigned to the different treatments. However, in a non-randomized observational study, investigators have no control over the treatment assignment, and therefore direct comparisons of outcomes from the treatment groups may be misleading. This difficulty may be partially avoided if information on measured covariates is incorporated into the study design (for example, through matched sampling) or into estimation of the treatment effect (for example, through stratification or covariance adjustment). Traditional methods of adjustment (matching, stratification and covariance adjustment) are often limited since they can only use a limited number of covariates for adjustment. However, propensity scores, which provide a scalar summary of the covariate information, do not have this limitation.

Formally, the propensity score[22] for an individual is the probability of being treated conditional on (or based only on) the individual's covariate values. Intuitively, the propensity score is a measure of the likelihood that a person would have been treated using only their covariate scores. Rosenbaum and Rubin[22] showed that the propensity score is a balancing score and can be used in observational studies to reduce bias through the adjustment methods mentioned above.

The three goals of this tutorial are: to present the formal definition of propensity scores with some theoretical findings; to illustrate common uses of the propensity score; and to present applied examples that illustrate applications of the propensity score. The Appendix includes SAS code used to perform some of the analyses presented. The tutorial will conclude with a discussion about areas of current and future research.

DEFINITION

With complete data, Rosenbaum and Rubin[22] introduced the propensity score for subject $i (i = 1, \ldots, N)$ as the conditional probability of assignment to a particular treatment $(Z_i = 1)$ versus control $(Z_i = 0)$ given a vector of observed covariates, x_i:

$$e(x_i) = \text{pr}(Z_i = 1 \mid X_i = x_i)$$

where it is assumed that, given the X's, the Z_i are independent:

$$\text{pr}(Z_1 = z_1, \ldots, Z = z_N \mid X_1 = x_1, \ldots, X_N = x_N) = \prod_{i=1}^{N} e(x_i)^{z_i}\{1 - e(x_i)\}^{1-z_i}.$$

The propensity score is the 'coarsest function' of the covariates that is a balancing score, where a balancing score, $b(X)$, is defined as 'a function of the observed covariates X such that the conditional distribution of X given $b(X)$ is the same for treated $(Z = 1)$ and control $(Z = 0)$ units'.[22] For a specific value of the propensity score, the difference between the treatment and control means for all units with that value of the propensity score is an unbiased estimate of the average treatment effect at that propensity score, if the treatment assignment is strongly ignorable, given the covariates. Thus, matching, stratification, or regression (covariance) adjustment on the propensity score tends to produce unbiased estimates of the treatment effects when treatment assignment is strongly ignorable. Treatment assignment is considered strongly ignorable if the treatment assignment, Z, and the response, Y, are known to be conditionally independent given the covariates, X (that is, when $\mathbf{Y} \perp \mathbf{Z} \mid \mathbf{X}$).

When covariates contain no missing data, the propensity score can be estimated using discriminant analysis or logistic regression. Both of these techniques lead to estimates of probabilities of treatment assignment conditional on observed covariates. Formally, the observed covariates are assumed to have a multivariate normal distribution (conditional on Z) when discriminant analysis is used, whereas this assumption is not needed for logistic regression.

A question that may arise from investigators who have not used propensity scores before is: 'Why must we estimate the probability that a subject receives a certain treatment since we know for certain which treatment was given?' An answer to this question is that if we use the probability that a subject would have been treated (that is, the propensity score) to adjust our estimate of the treatment effect, we can create a 'quasi-randomized' experiment. That is, if we find two subjects, one in the treated group and one in the control, with the same propensity score, then we could imagine that these two subjects were 'randomly' assigned to each group in the sense of being equally likely to be treated or control. In a controlled experiment, the randomization, which assigns pairs of individuals to the treated and control groups, is better than this because it does not depend on the investigator conditioning on a particular set of covariates; rather it applies to any set of observed or unobserved covariates. Although the results of using the propensity scores are conditional only on the observed covariates, if one has the ability to measure many of the covariates that are believed to be related to the treatment assignment, then one can be fairly confident that approximately unbiased estimates for the treatment effect can be obtained.

USES OF PROPENSITY SCORES

Currently in observational studies, propensity scores are used primarily to reduce bias and increase precision. The three most common techniques that use the propensity score are matching, stratification (also called subclassification) and regression adjustment. Each of these techniques is a way to make an adjustment for covariates prior to (matching and stratification) or while (stratification and regression adjustment) calculating the treatment effect. With all three techniques, the propensity score is calculated the same way, but once it is estimated it is applied differently. Propensity scores are useful for these techniques because by definition the propensity score is the conditional probability of treatment given the observed covariates $e(X) = \text{pr}(Z = 1 \mid X)$, which implies that Z and X are conditionally independent given $e(X)$. Thus, subjects in treatment and control groups with equal (or nearly equal) propensity scores will tend to have the same (or nearly the same) distributions on their background covariates.[6] Exact adjustments made using the propensity score will, on average, remove all of the bias in the background covariates. Therefore bias-removing adjustments can be made using the propensity scores rather than all of the background covariates individually.

MATCHING

Often investigators are confronted with studies where there are a limited number of treated patients and a larger (usually much larger) number of control patients. An example is a March of Dimes funded study examining the effects of post-term birth on neuropsychiatric, social and academic achievements among school-aged children (that is, 5–10 year old children). At the onset of the study, the investigators had a collection of over 9000 birth records (749 treated (post-term) babies and over 9000 potential control (term) babies), with prenatal and birth history

information. It was financially unfeasible for the investigators to collect outcome measurements on all potential subjects, so some form of sampling had to be performed.

Matching is a common technique used to select control subjects who are 'matched' with the treated subjects on background covariates that the investigator believes need to be controlled. Although the idea of finding matches seems straightforward, it is often difficult to find subjects who are similar (that is, can be matched) on all important covariates, even when there are only a few background covariates of interest. The investigators for the March of Dimes study had to confront this problem as they had more than ten variables on which they desired to match subjects.

Propensity score matching solves this problem by allowing an investigator to control for many background covariates simultaneously by matching on a single scalar variable. Prior to propensity score matching, a common matching technique was Mahalanobis metric matching using several background covariates.[23–27] Mahalanobis metric matching is employed by randomly ordering subjects, and then calculating the distance between the first treated subject and all controls, where the distance, $d(i, j)$, between a treated subject i and a control subject j is defined by the Mahalanobis distance:

$$d(i, j) = (u - v)^{\mathrm{T}} C^{-1} (u - v)$$

where u and v are values of the matching variables for treated subject i and control subject j, and C is the sample covariance matrix of the matching variables from the full set of control subjects. The control subject, j, with the minimum distance $d(i, j)$ is chosen as the match for treated subject i, and both subjects are removed from the pool. This process is repeated until matches are found for all treated subjects. One of the drawbacks of this technique is that it is difficult to find close matches when there are many covariates included in the model. As the number of dimensions on which the Mahalanobis distance is calculated increases, the average distance between observations increases as well. Propensity scores, on the other hand, can be calculated using many covariates, yet its score is still a scalar summary of the variables, and therefore matching is usually easy.

Rosenbaum and Rubin[6] outline three techniques for constructing a matched sample which use the propensity score: (i) nearest available matching on the estimated propensity score; (ii) Mahalanobis metric matching including the propensity score; and (iii) nearest available Mahalanobis metric matching within calipers defined by the propensity score.

Nearest available matching on the estimated propensity score. This method consists of randomly ordering the treated and control subjects, then selecting the first treated subject and finding the control subject with closest propensity score. Both subjects are then removed from consideration for matching and the next treated subject is selected. Rosenbaum and Rubin[6] suggest using the logit of the estimated propensity score to match (that is, $\hat{q}(X) = \log[(1 - \hat{e}(X))/\hat{e}(X)]$) because the distribution of $\hat{q}(X)$ is often approximately normal. Here, and in the following matching methods, recall the propensity score model may include many more covariates than employed in the Mahalanobis distance calculations.

Mahalanobis metric matching including the propensity score. This procedure is performed exactly as described above for Mahalanobis metric matching, with an additional covariate, the logit of the estimated propensity score ($\hat{q}(X)$) included with the other covariates in the calculation of the Mahalanobis distance. Rubin[24] showed that when the covariates have multivariate normal distributions and the treated and control groups have a common

covariance matrix, Mahalanobis metric matching is an equal per cent bias reducing (EPBR) technique, where the bias is the mean for the treated minus the mean for the control. In other words, the per cent bias reduced on all covariates is equal, and there are no covariates (or linear combinations of covariates) whose bias will increase due to matching.

Nearest available Mahalanobis metric matching within calipers defined by the propensity score. This method combines the previous two methods into one. The treated subjects are randomly ordered, and the first treated subject is selected. All control subjects within a preset amount (or caliper) of the treated subject's estimated propensity score ($\hat{e}(X)$) or estimated logit of the propensity score ($\hat{q}(X)$) are then selected, and Mahalanobis distances, based on a smaller number of covariates, are calculated between these subjects and the treated subject. The closest control subject and the treated subject are then removed from the pool, and the process is repeated. All remaining control subjects are available for the next matching with a treated subject. The size of the caliper is determined by the investigator. Cochran and Rubin[23] give advice on how large a caliper should be chosen based on the average of the variances of the covariates in the treated and control groups. Rosenbaum and Rubin[6] suggest that caliper size of a quarter of a standard deviation of the logit of the propensity score be used.

All three methods are useful techniques for reducing bias. Rosenbaum and Rubin[6] concluded that nearest available matching on the estimated propensity score was the easiest technique in terms of computational considerations. The second method, Mahalanobis metric matching including the propensity score, 'produced smaller standardized differences for individual variables but left a substantial difference along the propensity score'. They found the third method, nearest available Mahalanobis metric matching within calipers defined by the propensity score, to be the best technique among the three. It produced the best balance between the covariates in the treated and control groups, as well as the best balance of the covariates' squares and cross-products between the two groups. This third technique can be considered in the following way. By defining calipers based on the propensity score, the investigator is trying to create the 'quasi-randomized' experiment discussed above. Then, the use of Mahalanobis metric matching within calipers on a subset of the important covariates to choose subjects can be likened to blocking on important background variables in randomized controlled experiments. Rubin[28] also discusses this interpretation of Mahalanobis metric matching within calipers based on the propensity score.

APPLIED EXAMPLE: MARCH OF DIMES MATCHING

We now describe the steps taken in the March of Dimes study where the Mahalanobis metric matching within calipers based on the propensity score method was employed. First, we examined the distribution of 13 background covariates on which the investigators desired to match the term and post-term subjects. Table I contains descriptive statistics for these 13 covariates and the logit of the estimated propensity score, separately for the term and post-term groups. The first four columns contain the mean and standard deviation for each covariate and the logit of the estimated propensity score and the last two columns contain two statistics that are used to compare the groups. These statistics are the two-sample t-statistic and the standardized percentage difference. Based on these statistics, we see that there is moderate to large differences between the term and post-term groups on several covariates.

Table I. Group comparisons prior to matching

Variable	Post-term		Term		Comparisons	
	Mean	SD	Mean	SD	Two-sample t-statistic	Standardized difference in %†
	$N = 749$		$N = 9241$			
Sex of child	0·527	0·500	0·500	0·500	1·42	5·4
Parity	0·697	1·12	0·790	1·01	− 2·40*	− 8·7
Mother's age (years)	28·2	5·20	28·8	5·1	− 3·38**	− 12·7
Deliverey mode	1·28	0·455	1·23	0·431	2·75**	10·2
Hobel prenatal score	8·20	7·09	9·05	7·50	− 2·99**	− 11·6
Hobel Intrapartum score	10·09	8·62	7·41	7·46	9·37**	33·3
Child's age (months)	23·01	11·58	22·19	13·34	1·62	6·5
Child's birthweight (log grams)	8·20	0·143	8·11	0·149	15·58**	60·3
Mother's race (white = 1, non-white = 2)	1·19	0·488	1·22	0·539	− 1·77	− 6·7
Class (high = 3, low = 1)	1·628	0·778	1·650	0·759	− 0·79	− 3·0
Antepartum complications (yes/no)	0·729	0·445	0·699	0·459	1·71	6·5
Vaginal bleeding (yes/no)	0·128	0·335	0·124	0·329	0·36	1·4
Abnormal labour (yes/no)	0·453	0·498	0·354	0·478	5·42**	20·6
Logit of the propensity score	2·15	0·798	2·83	0·797	− 22·34**	− 60·0

* $0{\cdot}05 > p > 0{\cdot}01$

** $p < 0{\cdot}01$

† The standardized difference in % is the mean difference as a percentage of the average standard deviation: $100(\bar{x}_p - \bar{x}_t)/\sqrt{\{(s_p^2 + s_t^2)/2\}}$, where for each covariate $\bar{x}_p$ and $\bar{x}_t$ are the sample means in the post-term and term groups, respectively, and s_p^2 and s_t^2 are the corresponding sample variances

Two covariates with large initial differences between the term and post-term groups are the Hobel prenatal risk score and the Hobel intrapartum risk score.[29] Both of these have large two-sample t-statistics and standardized percentage differences. The Hobel risk scores were prenatal and intrapartum complications scales, respectively. These scores were calculated by determining whether or not certain risks were present for the women during the pregnancy and labour and then assigning specified weights to the presence of each risk factor. They were then calculated as the sum of these weights. For instance, for the Hobel prenatal risk score, a weight of 10 was given if the pregnant woman had a previous stillbirth, a weight of 5 was given if the pregnant woman was $\geqslant 35$ or $\leqslant 15$ years of age, and a weight of 5 was given if the pregnant woman had an abnormal glucose tolerance test. Thus, a woman who had these three risks, and no others, had a score of 20. In addition to covariates measured on the mother, there were two variables measured on the newborn infant, gender and date of birth (which is used to determine the subject's age at the time of the study). The goal of the matching was to reduce the differences between the term and post-term subjects on each of the covariates.

A propensity score model was estimated using discriminant analysis. In addition to the 13 covariates in Table I, 15 additional variables were included in this model. These 15 variables consisted of seven interactions terms and eight quadratic terms, which were calculated from the original 13 covariates, giving a total of 28 terms in the model. These interaction and quadratic terms were included based upon both statistical and scientific criteria (that is, certain interactions

Table II. Group comparisons after matching for variables used in Mahalanobis metric matching

Variable	Post-term		Term		Comparisons	
	Mean	SD	Mean	SD	Two-sample t-statistic*	Standardized difference in %
	$N = 749$		$N = 749$			
Sex†	0·527	0·500	0·527	0·500	0·00	0·0
Parity	0·697	1·12	0·629	0·997	1·24	6·4
Mother's age (years)	28·2	5·20	28·1	4·68	0·40	2·1
Delivery mode	1·28	0·455	1·28	0·452	0·01	0·0
Hobel prenatal score	8·20	7·09	7·63	6·53	1·62	8·4
Hobel intrapartum score	10·09	8·62	9·72	8·13	0·87	4·5
Child's age (months)	23·01	11·58	23·0	11·25	0·01	0·07
Child's birthweight (log grams)	8·20	0·143	8·20	0·129	0·82	4·4
Mother's race (white = 1, non-white = 2)	1·19	0·488	1·19	0·460	− 0·03	0·2
Class (high = 3, low = 1)	1·628	0·778	1·676	− 0·738	− 1·23	− 6·3
Antepartum complications (yes/no)	0·729	0·445	0·716	0·451	0·57	2·9
Vaginal bleeding (yes/no)	0·128	0·335	0·097	0·295	1·94	10
Abnormal labour (yes/no)	0·453	0·498	0·428	0·495	0·97	5
Logit of the propensity score	2·15	0·798	2·18	0·773	− 0·68	− 2·5

* p-values for all t-tests larger than 0·05
† Sex was exactly matched by design

and quadratic terms were included even if they were not found to be statistically significant). Recall, we do not use the propensity score model to make inferential statements concerning the term and post-term groups, rather, we use it to find propensity scores which are used to match term and post-term subjects and therefore create balance between the term and post-term groups. Thus, estimating a propensity score model with many terms does not create a problem.

Once the propensity scores were estimated, Mahalanobis metric matching was preformed as follows. The post-term subjects were randomly ordered, and the first post-term subject was selected. All term subjects within a determined caliper of the post-term subject's estimated logit of the propensity score ($\hat{q}(X)$) were then selected. The caliper chosen was equal to one-quarter of a standard deviation of the logit of the propensity score (0·20 from Table I). For example, if the first post-term subject had an estimated propensity score equal to 2·3 on the logit scale, then all term subjects with logit propensity scores between 2·1 and 2·5 would be selected as potential matches. The next step is to calculate Mahalanobis distances between these term subjects and the post-term subject. The closest term subject and the post-term subject are then removed from the pool, and the process is repeated. From the 28 terms in the propensity score model, the investigators chose eight covariates (the first eight variables of Table I) and the propensity score to be used in the Mahalanobis metric matching. These eight covariates were chosen because they were determined to be the most important variables for the final matching.

Propensity score matching using this method succeeded in removing most of the bias between the term and post-term groups. Table II contains descriptive statistics (means and standard deviations), two-sample t-statistics and standardized percentage differences for the original

Table III. Per cent reduction in bias for variables with initial standardized bias greater than 20 per cent

Variable	Initial bias	Bias after matching	Per cent reduction*
Hobel intrapartum risk score	2·687	0·377	85·9
Child's birthweight (log grams)	0·088	0·006	93·2
Abnormal labour (yes/no)	0·099	0·025	74·7
Logit of the propensity score	− 0·677	− 0·028	95·9

* Per cent reduction equals $100(1 - b_m/b_i)$ where b_m and b_i are post-term minus term differences in covariate means after matching and initially, respectively

post-term group ($N = 749$) and the matched term group ($N = 749$ matched term subjects out of the original $N = 9241$ term subjects). As can be seen, the matched sample had similar means for each of the 13 covariates included in the model. Table III shows the bias reduction for the four covariates with the largest initial bias: the Hobel Intrapartum Risk Score; child's birthweight; abnormal labour indicator, and the logit of the propensity score. As can be seen, each of these covariates had over 74 per cent bias reduction after matching.

The tables describe the results based on choosing the best available term match for each post-term subject. In addition to this, we generated a list of potential matches for each of the 749 post-term babies. For each post-term baby we provided the investigators with a list of 15 potential term matches. Based on this information, the investigators were able to identify matches for a subset of the post-term babies and then gather data on the matched pairs. Currently, the data on the matched pairs has been collected and analyses have begun to examine the hypothesis of interest, that is whether post-term birth is associated with neuropsychiatric, social and academic achievements among school-aged children (that is, 5–10 year old children).

In this example, the use of propensity scores proved useful. In particular, we were able to assess the fit of the propensity score model and compare the balance of background covariates prior to committing any resources (time or money) to collecting outcome data on the matched controls. Also, it is important to realize that, since these comparisons involve only covariates and not outcome variables, there is no chance of biasing results in favour of one treatment condition versus the other through the selection of matched controls.

STRATIFICATION

Stratification (sometimes referred to as subclassification) is also commonly used in observational studies to control for systematic differences between the control and treated groups. This technique consists of grouping subjects into strata determined by observed background characteristics. Once the strata are defined, treated and control subjects who are in the same stratum are compared directly. Many of the same problems occur in stratification as with matching when the number of covariates increases. Cochran[30] notes that as the number of covariates increases, the number of strata grows exponentially. For instance, if all covariates were dichotomous categorical variables, then there would be 2^k subclasses for k covariates. If k is large, then some strata might contain subjects from only the treated group, which would make it impossible to estimate a treatment effect in that stratum. Here again the propensity score is very useful. Because the propensity score is a scalar summary of all the observed background covariates, stratification on it alone can balance the distributions of the covariates in the treated and control groups without the exponential increase in number of strata.

Rosenbaum and Rubin[22] present theoretical results showing that perfect stratification based on the propensity score will produce strata where the average treatment effect within strata is an unbiased estimate of the true treatment effect. Again they assume that the treatment assignment is strongly ignorable. Rosenbaum and Rubin[5] state that Cochran's[31] result, which indicated that creating five strata removes 90 per cent of the bias due to the stratifying variable or covariate, holds for stratification based on the propensity score. They state that, in fact, stratification on the propensity score balances all k covariates that are used to estimate the propensity score, and often five strata based on the propensity score will remove over 90 per cent of the bias in each of these covariates.

The technique used for determining strata is straightforward. First, the propensity score is estimated by logistic regression or discriminant analysis. The investigator then must decide whether the stratum boundaries should be based on the values of the propensity score for both groups combined or in the treated or control group alone. Typically, in our work, we use the quintiles of the estimated propensity score from the combined group to determine the cut-offs for the different strata.

There are many examples in the recent literature of studies that have used propensity scores for stratification.[5,9,16,17,19–21] We now describe briefly some of these studies.

In Stone[20] investigators wished to compare outcomes on 747 patients with community-acquired pneumonia (CAP) who were either hospitalized ($n = 265$) or ambulatory ($n = 482$). Since patients were not randomized to be either hospitalized or ambulatory, propensity scores were estimated using classification tree techniques. Patients were then assigned to one of seven strata based on their estimated propensity score. The investigators found that there were imbalances between the two groups on 29 out of 44 baseline variables, and that after stratification on the propensity score only 13 of these remained significant at $p = 0{\cdot}05$. The investigators then estimated treatment effects using direct standardization methods of the stratum-specific means.

Fiebach *et al.*[9] used propensity scores to stratify patients who had received one of two possible treatments when they came to a hospital with uncomplicated chest pain. The two treatments were either admittance to a stepdown unit or admittance to a coronary care unit. Covariates used to estimate the propensity score included variables for the actual triage location and independent clinical predictors for an adverse event. These clinical predictors consisted of more than 50 clinical characteristics. A stepwise procedure was used to estimate the propensity score where covariates were entered into the model if they were significant at the 0·50 level in a stepwise discriminant analysis.

In Rosenbaum and Rubin[5] the authors wished to study the properties of the propensity score when used to stratify subjects in different treatment groups. In their example, the propensity score was the probability of receiving either coronary artery bypass surgery or medical therapy given 74 different covariates. These covariates consisted of haemodynamic, angiographic, laboratory and exercise test results. The investigators used a multi-stage procedure to find the best model for the propensity score. They found that using five strata based on the estimated propensity score was able to substantially reduce the bias in all 74 covariates simultaneously.

APPLIED EXAMPLE: FROM THE ACT STUDY

To illustrate further how to estimate and use the propensity score for stratification, we now present an applied example using data from the Active Management of Labor Trial (ACT).[32] The ACT trial is a randomized experiment to study the effects of active management of labour on the

Table IV. Comparison of covariates for subjects with and without epidural before and after propensity score stratification

	No epidural ($N = 775$) mean (sd)	Epidural ($N = 1003$) mean (sd)	F-statistics before stratification†	F-statistics after stratification‡
Pregnancy and labour characteristics				
Treated with active management of labour protocol	0·337 (0·47)	0·279 (0·45)	6·87**	0·20
Centimetres dilated at admission	3·95 (1·96)	2·79 (1·42)	208·01***	0·65
Artificially ruptured membranes (yes/no)	0·556 (0·50)	0·594 (0·49)	2·60	0·03
Gestational age (weeks)	39·9 (1·24)	40·2 (1·24)	22·28***	0·17
Infant birthweight (grams)	3374 (401)	3463 (416)	20·65***	0·20
Infant's gender (male = 1)	0·529 (0·50)	0·510 (0·50)	0·60	0·28
Initial rate of cervical dilation	58·3 (28·3)	42·9 (27·1)	135·20***	0·70
Maternal chronic hypertension	0·026 (0·16)	0·021 (0·14)	0·46	0·03
Maternal pregnancy induced hypertension (yes/no)	0·023 (0·15)	0·028 (0·16)	0·38	0·17
Maternal demographic/physical characteristics				
Maternal height (inches)	64·9 (2·8)	64·5 (2·6)	11·14**	0·10
Maternal pre-pregnant weight (pounds)	131·3 (21·6)	133·9 (22·9)	5·58*	0·07
Mother's age (years)	29·3 (5·1)	29·4 (5·3)	0·19	0·43
Insurance: private	0·857 (0·35)	0·882 (0·32)	2·55	2·75
public	0·101 (0·30)	0·084 (0·28)	1·51	0·54
Maternal race: white	0·677 (0·47)	0·735 (0·44)	7·01**	0·07
black	0·134 (0·34)	0·127 (0·33)	0·17	0·12
Hispanic	0·080 (0·27)	0·071 (0·26)	0·54	0·03

*$0{\cdot}05 > p > 0{\cdot}01$ **$0{\cdot}01 > p > 0{\cdot}001$ ***$0{\cdot}001 > p$

† F-statistic = square of two-sample t-statistic

‡ F-statistic for main effect of epidural use after adjusting for propensity score quintile

probability of having a Caesarean section. There were two components to this trial: a baseline component and a randomized component. In addition to the original study questions, the investigators were interested in determining whether the use of epidural anaesthesia was associated with Caesarean section in nulliparous women. To study this question, they wished to examine all eligible women from the baseline and randomized components of the trial. Propensity scores were used in these analyses since women were not randomly assigned to receive the treatment (an epidural). In this report we include 1778 women in the analyses, of these 1003 (56·4 per cent) had received an epidural. The investigators identified 15 variables (Table IV) which they felt may be imbalanced between the women who received an epidural and those who did not. Table IV shows the covariate imbalance before and after stratification based on the quintiles of the propensity score. Ten of the covariates were included in the final propensity score model used for stratification. The initial imbalance was measured by calculating F-statistics (squares of two-sample t-statistics) comparing the epidural and no epidural groups.

Propensity scores were estimated for each woman using logistic regression. Two of the covariates in this model, mother's insurance and mother's race, were transformed into dummy variables for the logistic regression model. Women were then separated into quintiles defined by

Table V. Comparison of quintile means for variable centimetres at admission

		N	Centimetres dilated at admission Mean (SD)
Overall	No epidural	775	3·95 (1·96)
	Epidural	1003	2·79 (1·42)
After stratification into quintiles based on propensity scores			
Quintile 1	No epidural	55	1·93 (1·02)
	Epidural	263	1·90 (1·03)
Quintile 2	No epidural	83	2·55 (1·00)
	Epidural	236	2·62 (1·11)
Quintile 3	No epidural	126	3·00 (1·28)
	Epidural	193	3·05 (1·19)
Quintile 4	No epidural	157	3·54 (1·32)
	Epidural	162	3·61 (1·42)
Quintile 5	No epidural	268	5·40 (1·78)
	Epidural	50	4·68 (1·19)

their propensity scores. We then compared the epidural/no epidural groups on their 15 covariates, after adjusting for their propensity quintile. This was done using a two-way analysis of variance model which included main effects for propensity score quintile (coded as a class variable with 4 degrees of freedom) and epidural use (coded as yes/no). We compared the F-statistic for epidural use after adjustment for propensity score quintile with the F-statistic for epidural use prior to adjustment for propensity score quintile to determine whether balance was achieved after stratification based on the propensity score. We also examined the two-way interaction of quintile and epidural use. We found that the eight covariates which were significantly different between the two groups prior to stratification, were all non-significantly different after adjustment for propensity score quintile (see Table IV). Among the interaction terms, only one was significant for the variable centimetres dilated at first exam ($F = 3{\cdot}16$, $p = 0{\cdot}013$). We further examined this variable by presenting the quintile means for the epidural and no epidural group (Table V and Figure 1). As can be seen by this table and figure, in the first four quintiles the mean centimetres at admission are very close, whereas in the 5th quintile the two groups are still somewhat separated. As can be seen in Figure 1, the means for the epidural/no epidural groups cross, and this explains the significant interaction between epidural use and propensity score quintile. Nevertheless, as can be seen both in the table and figure, the two groups are more similar within each propensity score quintile than they were before stratification. We have included SAS code which was used to estimate the propensity scores and perform the analyses presented here in the Appendix.

The investigators had several options for how they would estimate the effects of epidural use on the rate of Caesarean section use employing propensity scores. One method would be to estimate the treatment effects separately within each quintile defined by the propensity score and then combine the quintile estimates into an overall estimate of the treatment effect. An alternative method would be to perform a multiple logistic regression with use of Caesarean as the outcome and epidural use as the independent variable. In this model the propensity score could be

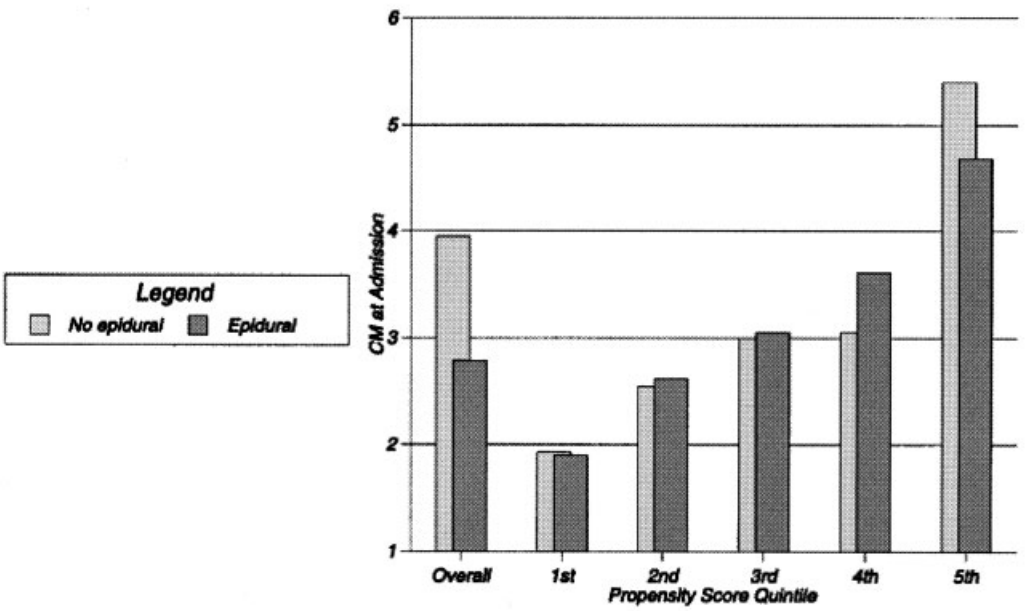


Figure 1. Comparison of quintile means for variable centimetres (cm) dilated at admission

included either as its raw score or the propensity score quintile itself could be included. In addition to the propensity score, a subset of the other covariates could be included (such as centimetres at admission). One advantage of this method is that the final model to estimate the treatment effect contains fewer covariates than if all the covariates in the propensity score model had been used. Therefore traditional diagnostics can be more easily employed to determine the fit of this final model. This was the method employed in this particular application, and the results indicated that after adjustment for propensity score (either the raw score or quintile) and a subset of important covariates, the rate of use of Caesarean sections in the epidural group was still significantly higher than the no epidural group (odds ratio = 3·7 with a 95 per cent confidence interval from 2·4 to 5·7).[21]

REGRESSION (COVARIANCE) ADJUSTMENT

Propensity scores can also be used in regression (covariance) adjustment. In regression adjustment, the treatment effect, τ, is estimated as

$$\hat{\tau} = (\bar{Y}_t - \bar{Y}_c) - \beta(\bar{X}_t - \bar{X}_c)$$

where the t and c indicate treatment and control groups. The effect of the background covariates is adjusted for by subtracting out the second term on the right hand side of the above equation, where β is an estimate of the regression of the responses for the treated and control groups on the background covariates. For the reasons stated above, the propensity score is a useful variable in regression adjustments, since one only has to find the regression of the responses on the propensity scores in the treated and control groups and use this to adjust the final estimate of the

treatment effect. Roseman[33] finds that if the response surfaces in the treatment and control groups are parallel and either linear or non-linear, then the regression adjustment using the propensity scores reduces the bias in the estimate of the treatment effect. In addition, if one stratifies and then uses regression adjustment within the strata, then this estimator of estimated treatment effect appears to be a more efficient estimator than one based on matching alone.

Another approach to regression adjustment is to use a large set of background covariates to estimate the propensity score and then take a subset of these covariates and the propensity score and use them in the regression adjustment. This is the analysis performed above by the ACT investigators. This method is also analogous to performing Mahalanobis metric matching within calipers on a subset of the important covariates including the propensity score as discussed above.

As with matching and subclassification, there are examples in the literature where regression adjustments were used with propensity scores.[10,14,15] Here is a brief description of two of them.

In Berk and Newton,[14] investigators wished to determine whether new spousal violence was influenced by whether men were either arrested or not arrested for wife-battery. Here the propensity score was the probability of being arrested and 14 covariates were used to estimate it. Some of these covariates included the suspect's age, whether the victim was injured, whether the suspect was drinking, and whether the victim had called the police. The investigators then 'regressed whether or not there was new spousal violence during the follow-up (by the same offender against the same victim) on the propensity scores, *separately for the arrested and not arrested group*'. They found that the arrested and not arrested groups had similar intercepts but different slopes. For the arrested group, the slope was near zero, but for the not arrested group the slope was very steep. This seemed to suggest that subjects who had a high propensity to be arrested, but were not arrested, were more likely to commit violence during the follow-up period.

In Muller *et al.*,[10] investigators studied the effects of digoxin in mortality rates in patients after myocardial infarction. Here the non-randomized treatment was digoxin. The investigators calculated an 'imbalance risk score', which appears to be a propensity score, based on 19 covariates. The covariates in this model included the subject's heart rate, age, and whether the subject had beta-blockers in the previous three weeks. Cox proportional-hazards regression was used to determine the association between digoxin therapy and survival 'taking into account the effects of baseline prognostic factors'. It appears that they 'took into account' the baseline differences by including a propensity score as a covariate in their model in order to adjust their final treatment effect.

One question which may arise when using regression adjustment with propensity scores is whether there is any gain in using the propensity score rather than performing a regression adjustment with all of the covariates used to estimate the propensity score included in the model. Rosenbaum and Rubin[25] showed that the 'point estimate of the treatment effect from an analysis of covariance adjustment for multivariate X is ... equal to the estimate obtained from a univariate covariance adjustment for the sample linear discriminant based on X, whenever the same sample covariance matrix is used for both the covariance adjustment and the discriminant analysis'. Thus, the results from both methods should lead to the same conclusions. However, one advantage to performing the two-step procedure is that one can fit a very complicated propensity score model with interactions and higher order terms first. Since the goal of this propensity score model is to obtain the best estimated probability of treatment assignment, one is not concerned with over-parameterizing this model. Then when the model for estimating the treatment effect is estimated the investigator can include only a subset of the most important variables and the

propensity score in the model. This smaller model may allow the investigator to perform diagnostic checks on the fit of the model more reliably than if there were many covariates included in the model.

In general, covariance adjustment should be performed with caution. Rubin[25] showed that covariance adjustment may in fact increase the expected squared bias if the covariance matrices in the treated and untreated groups are unequal (that is, if the discriminant is *not* a monotone function of the propensity score). Another difficulty arises when the variance in the treated and untreated groups are very different (that is, the untreated group variance is much larger than the treated groups variance). Under these circumstances, one may consider using propensity score methods for matching or subclassification, rather than using covariance adjustment.

DISCUSSION AND CURRENT RESEARCH

In all the examples stated above, except for Rosenbaum and Rubin,[5] there was no mention of how missing values on covariates were handled when estimating propensity scores. This is an important issue in most real data applications. For instance, in the ACT example presented above, over 100 subjects would have been excluded from the final analyses based on the fact that they were missing covariates included in the propensity score models. The March of Dimes Study also had covariates with missing data into the models. Currently, methods are being developed to handle this problem.[34,35] which allow for different missing-data mechanisms and use the EM[36] or ECM[37] algorithms to estimate propensity scores.

A second area of current research is using propensity scores to estimate treatment effects in clinical trials where subjects drop out prior to the trials completion.[38] Here, propensity scores are estimated as the probability that an individual will complete the trial conditional on their baseline and early outcomes. This work also uses the EM and ECM algorithms to estimate propensity scores with missing data.

Propensity scores are being widely used in statistical analyses, particularly in the area of applied medicine. Their use should only increase as the cost for randomized clinical trials rises and more investigators turn to observational studies as a means of performing less expensive research. The propensity score methodology appears to produce the greatest benefits when it can be incorporated into the design stages of studies (through matching or stratification). These benefits include providing more precise estimates of the true treatment effects as well as saving time and money. This saving results from being able to avoid recruitment of subjects who may not be appropriate for particular studies. Finally, it is important to note that we are not advocating the use of only propensity scores in analyses of observational studies, rather we are encouraging the use of propensity scores in addition to traditional methods of analysis. The propensity score should be thought of as an additional tool available to the investigators as they try to estimate the effects of treatments in studies.

APPENDIX

The following is a series of SAS code that will estimate propensity scores using the logistic regression procedure in SAS. We first perform a series of t-tests to determine the initial level of bias between the two groups. Then we calculate propensity scores using stepwise logistic regression. Next, propensity score quintiles are created and then we examine whether the groups are balanced after adjustment for the propensity score quintiles.

```
* This will perform a series of ttests to determine what the initial difference between the
* treated and control groups are. Here the variable epidural is the treatment indicator in this
* model.

proc ttest data = matchset; class epidural;
var amladmit cm 1 arom gestage birthwt gender rate chyper phyper
    height weight momage insprvt momw momb;

* This performs a stepwise logistic regression to estimate propensity scores for each subject.
* The variable pr is the propensity score The variable epidural is the treatment indicator in
* this model.

proc logistic data = matchset nosimple;
model epidural = amladmit cm 1 arom gestage birthwt gender rate chyper phyper
    height weight momage insprvt momw momb/selection = stepwise;
output out = preds pred = pr;

* This takes the propensity score and creates quintiles based on the estimated propensity
* score;

proc rank groups = 5 out = r;
ranks rnks;
var pr;
data a; set r; quintile = rnks + 1;

* This will show the breakdown of subjects by treatment (here epidural) and propensity
  score quintile;
proc freq; tables quintile*epidural;

* This will perform the 2-way anovas to determine whether the propensity score quintiles
* removed the initial bias found by the t-tests above.
proc glm;
    class quintile;
    model amladmit cm 1 arom gestage birthwt gender rate chyper phyper
    height weight momage insprvt momw momb = quintile epidural quintile*epidural;
```

ACKNOWLEDGEMENTS

The author wishes to thank Dr. Ellice Lieberman and the investigators from the Active Management of Labor Trial (National Institute of Child Health and Human Development grant no. R01-HD26813) for use of their data. The author also wishes to thank Dr. Curtis Deutch and the March of Dimes Birth Defects Foundation Social and Behavioral Science Research grant for use of their data. The author also wishes to acknowledge the support of his wife Carey and daughters Lucy, Serena and Sophia in this work.

REFERENCES

1. Rubin, D. B. and Thomas, N. 'Matching using estimated propensity scores: Relating theory to practice', *Biometrics*, **52**, 249–264 (1996).
2. Bloch, D. A. and Segal, M. R. 'Empirical comparison of approaches to forming strata: using classification trees to adjust for covariates', *Journal of the American Statistical Association*, **84**, 897–905 (1989).
3. Ciampi, A., Hogg, S. A., McKinney, S. and Thiffault, J. 'RECPAM: a computer program for recursive partition and amalgamation for censored survival data and other situation frequently occurring in biostatistics. I. Methods and program features', *Computer Methods and Programs in Biomedicine*, **26**, 239–256 (1988).
4. Rosenbaum, P. R. 'Conditional permutation tests and the propensity score in observational studies', *Journal of the American Statistical Association*, **79**, 565–574 (1984).

5. Rosenbaum, P. R. and Rubin, D. B. 'Reducing bias in observational studies using subclassification on the propensity score', *Journal of the American Statistical Association*, **79**, 516–524 (1984).
6. Rosenbaum, P. R. and Rubin, D. B. 'Constructing a control group using multivariate matched sampling methods that incorporate the propensity score', *American Statistician*, **39**, 33–38 (1985).
7. Lavori, P. W. and Keller, M. B. 'Improving the aggregate performance of psychiatric diagnostic methods when not all subjects receive the standard test', *Statistics in Medicine*, **7**, 723–737 (1988).
8. Cook, E. F. and Goldman, L. 'Asymmetric stratification: An outline for an efficient method for controlling confounding in cohort studies', *American Journal of Epidemiology*, **127**, 626–639 (1988).
9. Fiebach, N. H., Cook, E. F., Lee, T. H., Brand, D. A., Rouan, G. W., Weisberg, M. and Goldman, L. 'Outcomes in patients with myocardial infarction who are initially admitted to stepdown units: data from the multicenter chest pain study', *American Journal of Medicine*, **89**, 15–20 (1990).
10. Muller, J. E., Turi, Z. G., Stone, P. H., Rude, R. E., Raabe, D. S., Jaffe, A. S., Gold, H. K., Gustafson, N., Poole, W. K., Passamani, E., Smith, T. W., Braunwald, E. and The MILIS Study Group. 'Digoxin therapy and mortality after myocardial infarction: experience in the MILIS Study', *New England Journal of Medicine*, **314**, 265–271 (1986).
11. Myers, W. O., Gersh, B. J., Fisher, L. D., Kosinski, A. S., Mock, M. B., Holmes, D. R., Schaff, H. V., Gillispie, S., Ryan, T. J., Kaiser, G. C. and other CASS Investigators. 'Time to first new myocardial infarction in patients with mild angina and three-vessel disease comparing medicine and early surgery: A Cass registry study of survival', *Annals of Thoracic Surgery*, **43**, 599–612 (1987).
12. Myers, W. O., Gersh, B. J., Fisher, L. D., Kosinski, A. S., Mock, M. B., Holmes, D. R., Schaff, H. V., Gillispie, S., Ryan, T. J., Kaiser, G. C. and other CASS Investigators. 'Multiple versus early surgical therapy in patients with triple-vessel disease and mild angina pectoris: A Cass Registry Study of Survival', *Annals of Thoracic Surgery*, **44**, 471–486 (1987).
13. Myers, W. O., Schaff, H. V., Gersh, B. J., Fisher, L. D., Kosinski, A. S., Mock, M. B., Holmes, D. R., Ryan, T. J., Kaiser, G. C. and CASS Investigators. 'Improved survival of surgically treated patients with triple vessel coronary artery disease and severe angina pectoris', *Journal of Thoracic and Cardiovascular Surgery*, **97**, 487–495 (1989).
14. Berk, R. A. and Newton, P. J. 'Does arrest really deter wife battery? An effort to replicate the findings of the Minneapolis Spouse Abuse Experiment', *American Sociological Review*, **50**, 253–262 (1985).
15. Berk, R. A., Newton, P. J. and Berk, S. F. 'What a difference a day makes: an empirical study of the impact of shelters for battered women', *Journal of Marriage and the Family*, **48**, 481–490 (1986).
16. Czajka, J. L., Hirabayashi, S. M., Little, R. J. A. and Rubin, D. B. 'Projecting from advance data using propensity modeling: an application to income and tax statistics', *Journal of Business and Economic Statistics*, **10**, 117–131 (1992).
17. Hoffer, T., Greeley, A. M. and Coleman, J. S. 'Achievement growth in public and Catholic schools', *Sociology of Education*, **58**, 74–97 (1985).
18. Lavori, P. W. 'Clinical trials in psychiatry: should protocol deviation censor patient data?', *Neuropsychopharmacology*, **6**, (1), 39–48 (1992).
19. Lavori, P. W., Keller, M. B. and Endicott, J. 'Improving the validity of Fh-Rdc diagnosis of major affective disorder in uninterviewed relatives in family studies: a model based approach', *Journal of Psychiatric Research*, **22**, 249–259 (1988).
20. Stone, R. A., Obrosky, S., Singer, D. E., Kapoor, W. N. and Fine, M. J. 'Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired pneumonia', *Medical Care*, **33**, AS56–AS66 (1995).
21. Lieberman, E., Lang, J. M., Cohen, A., D'Agostino, Jr. R., Datta, S. and Frigoletto, Jr., F. D., 'Association of epidural analgesia with caesareans in nulliparous women', *Obstetrics and Gynecology*, **88**, 993–1000 (1996).
22. Rosenbaum, P. R. and Rubin, D. B. 'The central role of the propensity score in observational studies for causal effects', *Biometrika*, **70**, 41–55 (1983).
23. Cochran, W. G. and Rubin, D. B. 'Controlling bias in observational studies: a review', *Sankya, Series A*, **35**, 417–446 (1973).
24. Rubin, D. B. 'Matching methods that are equal percent bias reducing: some examples', *Biometrics*, **32**, 109–120 (1976).
25. Rubin, D. B. 'Using multivariate matched sampling and regression adjustment to control bias in observational studies', *Journal of the American Statistical Association*, **74**, 318–324 (1979).

26. Rubin, D. B. 'Bias reduction using Mahalanobis metric matching', *Biometrics*, **36**, 293–298 (1980).
27. Carpenter, R. G. 'Matching when covariables are normally distributed', *Biometrika*, **64**, 299–307 (1977).
28. Rubin, D. B. 'Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism', *Biometrics*, **47**, 1213–1234 (1991).
29. Hobel, C. J., Youkeles, L. and Forsythe, A. 'Prenatal and intrapartum high-risk screening. II Risk factors reassessed', *American Journal of Obstetrics and Gynecology*, **135**, 1051–1056 (1979).
30. Cochran, W. G. 'The planning of observational studies of human populations', *Journal of the Royal Statistical Society, Series A*, **128**, 234–255 (1965).
31. Cochran, W. G. 'The effectiveness of adjustment by subclassification in removing bias in observational studies', *Biometrics*, **24**, 205–213 (1968).
32. Frigoletto, F. D., Lieberman, E., Lang, J. M., Cohen, A. P., Barss, V., Ringer, S. A. and Datta, S. 'A clinical trial of active management of labor', *New England Journal of Medicine*, **333**, 745–750 (1995).
33. Roseman, L. 'Using regression and subclassification on the propensity score to control bias in observational studies', unpublished report, Harvard University, 1994.
34. D'Agostino, R. B. Jr. 'Estimating propensity scores when covariates have either ignorable or nonignorable missing values', Ph.D. thesis, Harvard University, 1994.
35. D'Agostino, R. B. Jr. and Rubin, D. B. 'Estimating and using propensity scores with partially missing data', submitted to *Journal of the American Statistical Association*, (1977).
36. Dempster A. P., Laird N. M. and Rubin D. B. 'Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)' *Journal of the Royal Statistical Society, Series B*, **39**, 1–38 (1977).
37. Meng, X. L. and Rubin, D. B. 'Maximum likelihood estimation via the ECM algorithm: A general framework', *Biometrika*, **80**, 267–278 (1993).
38. Dawson, R. and D'Agostino Jr., R. B. 'Propensity-based non-ignorable models for drop-out in clinical trials', submitted to *Biometrics* (1996).

1.3 Agreement Statistics

TUTORIAL IN BIOSTATISTICS
Kappa coefficients in medical research

Helena Chmura Kraemer[1,*,†], Vyjeyanthi S. Periyakoil[2] and Art Noda[1]

[1]*Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, California, U.S.A.*
[2]*VA Palo Alto Health Care System, Palo Alto, CA, U.S.A.*

SUMMARY

Kappa coefficients are measures of correlation between categorical variables often used as reliability or validity coefficients. We recapitulate development and definitions of the K (categories) by M (ratings) kappas ($K \times M$), discuss what they are well- or ill-designed to do, and summarize where kappas now stand with regard to their application in medical research. The $2 \times M$ ($M \geqslant 2$) intraclass kappa seems the ideal measure of binary reliability; a 2×2 weighted kappa is an excellent choice, though not a unique one, as a validity measure. For both the intraclass and weighted kappas, we address continuing problems with kappas. There are serious problems with using the $K \times M$ intraclass ($K > 2$) or the various $K \times M$ weighted kappas for $K > 2$ or $M > 2$ in any context, either because they convey incomplete and possibly misleading information, or because other approaches are preferable to their use. We illustrate the use of the recommended kappas with applications in medical research. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: kappa; reliability; validity; consensus

1. INTRODUCTION

'Many human endeavors have been cursed with repeated failures before final success is achieved. The scaling of Mount Everest is one example. The discovery of the Northwest Passage is a second. The derivation of a correct standard error for kappa is a third'. This wry comment by Fleiss *et al.* in 1979 [1] continues to characterize the situation with regard to the kappas coefficients up to the year 2001, including not only derivation of correct standard errors, but also the formulation, interpretation and application of kappas.

*Correspondence to: Helena Chmura Kraemer, Department of Psychiatry and Behavioral Sciences, MC 5717, Stanford University, Stanford, CA 94305, U.S.A.
†E-mail: hck@leland.stanford.edu

Contract/grant sponsor: National Institute of Mental Health; contract/grant number: MH40041
Contract/grant sponsor: National Institute of Aging; contract/grant number: AG17824
Contract/grant sponsor: Department of Veterans Affairs Sierra-Pacific MIRECC
Contract/grant sponsor: Medical Research Service of the Department of Veterans Affairs

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
 ISBN: 0-470-02365-1

The various kappa coefficients are measures of association or correlation between variables measured at the categorical level. The first formal introductions of kappa were those, more than 40 years ago, by Scott [2] and Cohen [3]. Since then, the types of research questions in medical research that are well addressed with kappas (for example, reliability and validity of diagnosis, risk factor estimation) abound, and such areas of research have become of ever growing interest and importance [4]. Not surprisingly, numerous papers both using and criticizing the various forms of kappas have appeared in the statistical literature, as well as in the psychology, education, epidemiology, psychiatry and other medical literature. It is thus appropriate, despite the many existing 'revisits' of kappas [5–15], to take stock of what kappas are, what they are well-designed or ill-designed to do, and to bring up to date where kappas stand with regard to their applications in medical research.

To set the stage for discussion let us consider five major issues concerning kappas that are often forgotten or misinterpreted in the literature:

1. *Kappa has meaning beyond percentage agreement corrected for chance* (*PACC*). Sir Alexander Fleming in 1928 discovered penicillin by noticing that bacteria failed to grow on a mouldy Petri dish. However, in summarizing current knowledge of penicillin and its uses, a mouldy Petri dish is at most a historical curiosity, not of current relevance to knowledge about penicillin. In much the same way, Jacob Cohen discovered kappa by noticing that this statistic represented percentage agreement between categories corrected for chance (PACC). Since then, there has also been much expansion and refinement of our knowledge about kappa, its meaning and its use. Whether to use or not use kappa has very little to do with its relationship to PACC. With regard to kappa, that relationship is a historical curiosity. Just as some scientists study moulds, and others bacteria, to whom penicillin is a side issue, there are scientists specifically interested in percentage agreement. To them whether rescaling it to a kappa is appropriate to its understanding and use is a side issue [16–20]. Consequently there are now two separate and distinct lines of inquiry, sharing historical roots, one concerning use and interpretation of percentage agreement that will not be addressed here, and that concerning use and interpretation of kappa which is here the focus.
2. *Kappas were designed to measure correlation between nominal, not ordinal, measures*. While the kappas that emerged from consideration of agreement between non-ordered categories can be extended to ordinal measures [21–23], there are better alternatives to kappas for ordered categories. Technically, one can certainly compute kappas with ordered categories, for example, certain, probable, possible and doubtful diagnosis of multiple sclerosis [24], and the documentation of many statistical computer programs (for example, SAS) seem to support this approach, but the interpretation of the results can be misleading. In all that follows, the measures to be considered will be strictly nominal, not ordered categories.
3. *Even restricted to non-ordered categories, kappas are meant to be used, not only as descriptive statistics, but as a basis of statistical inference*. RBI or batting averages in baseball are purely descriptive statistics, not meant to be used as a basis of statistical inference. Once one understands how each is computed, it is a matter of personal preference and subjective judgement which statistic would be preferable in evaluating the performance of batters. In contrast, means, variance, correlation coefficients etc., as they are used in medical research, are descriptive statistics of what is seen in a particular

sample, but are also meant to estimate certain clinically meaningful population characteristics, and to be used as a basis of inference from the sample to its population. To be of value to medical research, kappas must do likewise.

Nevertheless, presentations of kappas often do not define any population or any parameter of the population that sample kappas are meant to estimate, and treat kappas purely as descriptive statistics [7]. Then discussions of bias, standard error, or any other such statistical inference procedures from sample to population are compromised. Many of the criticisms of kappas have been based on subjective opinions as to whether kappas are 'fair to the raters' or 'large enough', behave 'as they should', or accord with some personal preference as to what 'chance' means [7, 13, 25, 26]. These kinds of discussions of subjective preferences are appropriate to discussing RBI versus batting average, but not to estimation of a well-defined parameter in a population. We would urge that the sequence of events leading to use of a kappa coefficient should be: (i) to start with an important problem in medical research; (ii) to define the population and the parameter that the problem connotes; (iii) to discuss how (or whether) sample kappa might estimate that parameter, and (iv) to derive its statistical properties in that population. When this procedure is followed, it becomes clear that there is not one kappa coefficient, but many, and that which kappa coefficient is used in which situation is of importance. Moreover, there are many situations in which kappa can be used, but probably should not be.

4. *In using kappas as a basis of statistical inference, whether or not kappas are consistent with random decision making is usually of minimal importance*. Tests of the null hypothesis of randomness (for example, chi-square contingency table analyses) are well established and do not require kappa coefficients for implementation. Kappas are designed as effect sizes indicating the degree or strength of association. Thus bias of the sample kappas (relative to their population values), their standard errors (in non-random conditions), computation of confidence intervals, tests of homogeneity etc. are the statistical issues of importance [27–30]. However, because of overemphasis on testing null hypotheses of randomness, much of the kappa literature that deals with statistical inference focuses not on kappa as an effect size, but on testing whether kappas are random or not. In this discussion no particular emphasis will be placed on the properties of kappas under the assumption of randomness.
5. *The use of kappas in statistical inference does not depend on any distributional assumptions on the process underlying the generation of the classifications*. However, many presentations impose such restricting assumptions on the distributions of $\mathbf{p}_i$ that may not well represent what is actually occurring in the population.

   The population model for a nominal rating is as follows. Patients in a population are indexed by i, $i = 1, 2, 3, \ldots$. A single rating of a patient is a classification of patient i into one of $K (K > 1)$ mutually exclusive and exhaustive non-ordered categories and is represented by a K-dimensional vector $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iK})$, where $X_{ij} = 1$, if patient i is classified into category j, and all other entries equal 0. For each patient, there might be M $(M > 1)$ such ratings, each done blinded to all the others. Thus any correlation between the ratings arises from correlation within the patients and not because of the influence of one rater or rating on another. The probability that patient i $(i = 1, 2, \ldots)$ is classified into category j $(j = 1, 2, \ldots, K)$ is denoted p_{ij}, and $\mathbf{p}_i$ is the K-dimensional vector $(p_{i1}, p_{i2}, \ldots, p_{iK})$ with non-negative entries summing to 1. In a

particular population of which patient i is a member, $\mathbf{p}_i$ has some, usually unknown, distribution over the $K-1$ dimensional unit cube.

For example, when there are two categories ($K=2$), for example, diagnosis of disease positive or negative, one common assumption is that the probability that a patient actually has the disease is π, and that if s/he has the disease, there is a fixed probability of a positive diagnosis ($X_{i1}=1$), the *sensitivity* (Se) of the diagnosis ($p_{i1}=\mathrm{Se}$); if s/he does not have the disease ($X_{i2}=2$), a fixed probability of a negative diagnosis, the *specificity* (Sp) of the diagnosis ($1-p_{i1}=\mathrm{Sp}$). This limits the distribution of p_{i1} to two points, Se and $1-\mathrm{Sp}$ ($p_{i2}=1-p_{i1}$): the 'sensitivity/specificity model' [31].

In the same situation, another model suggested has been the 'know/guess' model [25, 32, 33]. In this case, it is assumed that with a certain probability, π_1, a patient will be known with certainty to have the disease ($p_{i1}=1$); with a certain probability, π_0, a patient will be known with certainty not to have the disease ($p_{i1}=0$). For these patients, there is no probability of classification error. Finally, with the remaining probability, $1-\pi_1-\pi_0$, the diagnosis will be guessed with probability $p_{i1}=\alpha$. This limits the distribution of p_{i1} to 3 points $(1,\alpha,0)$.

One can check the fit of any such model by obtaining multiple blinded replicate diagnoses per patient. For these two models, three blinded diagnoses per patient would be required to estimate the three parameters in each model, $(\pi,\mathrm{Se},\mathrm{Sp})$ or (π_1,π_0,α), and at least one additional diagnosis per patient to test the fit of the model. In practice, it is hard to obtain four or more diagnoses per patient for a large enough sample size for adequate power, but in the rare cases where this has been done, such restrictive models are often shown to fit the data poorly [34]. If inferences are based on such limiting distributional assumptions that do not hold in the population, no matter how reasonable those assumptions might seem, or how much they simplify the mathematics, the conclusions drawn on that basis may be misleading. Kappas are based on no such limiting assumptions. Such models merely represent special cases often useful for illustrating certain properties of kappa, or for disproving certain general statements regarding kappa, as they here will be.

2. ASSESSMENT OF RELIABILITY OF NOMINAL DATA: THE INTRACLASS KAPPA

The reliability of a measure, as technically defined, is the ratio of the variance of the 'true' scores to that of the observed scores, where the 'true' score is the mean over independent replications of the measure [35, 36]. Since the reliability of a measure, so defined, indicates how reproducible that measure will be, how attenuated correlations against that measure will be, what loss of power of statistical tests use of that measure will cause, as well as how much error will be introduced into clinical decision making based on that measure [37], this is an important component of the quality of a measure both for research and clinical use. Since one cannot have a valid measure unless the measure has some degree of reliability, demonstration of reliability is viewed as a necessary first step to establishing the quality of a measure [14, 38].

The simplest way to estimate the reliability of a measure is to obtain a representative sample of N patients from the population to which results are to be generalized. (The same measure

may have different reliabilities in different populations.) Then M ratings are sampled from the finite or infinite population of ratings/raters to which results are to be generalized, each obtained blinded to every other. Thus the ratings might be M ratings by the same pathologist of tissue slides presented over a period of time in a way that ensures blindness: *intra-observer reliability*. The ratings might be diagnoses by M randomly selected clinicians from a pool of clinicians all observing the patient at one point in time: *inter-observer reliability*. The ratings might be observations by randomly selected observers from a pool of observers, each observing the patient at one of M randomly selected time points over a span of time during which the characteristic of the patient being rated is unlikely to change: *test–retest reliability*. Clearly there are many different types of reliability depending on when, by whom, and how the multiple blinded ratings for each patient are generated. What all these problems have in common is that because of the way ratings are generated, the M successive ratings per patient are 'interchangeable', that is, the process underlying the M successive ratings per patient has the same underlying distribution of $\mathbf{p}_i$, whatever that distribution might be [39].

2.1. The 2×2 intraclass kappa

The simplest and most common reliability assessment with nominal data is that of two ratings ($M = 2$), with two categories ($K = 2$). In that case, we can focus on the X_{i1} since $X_{i2} = 1 - X_{i1}$ and on p_{i1}, since $p_{i2} = 1 - p_{i1}$. Then $E(X_{i1}) = p_{i1}$, the 'true score' for patient i, $E(p_{i1}) = P, \text{variance}(p_{i1}) = \sigma_p^2$. Thus by the classical definition of reliability, the reliability of X is $\text{variance}(p_{i1})/\text{variance}(X_{i1}) = \sigma_p^2/PP'$, where $P' = 1 - P$.

This intraclass kappa, κ, may also be expressed as

$$\kappa = (p_0 - p_c)/(1 - p_c)$$

where p_0 is the probability of agreement, and $p_c = P^2 + P'^2$, that is, the PACC, for this has been shown to equal σ_p^2/PP' [31]. So accustomed are researchers to estimating the reliability of ordinal or interval level measures with a product-moment, intraclass or rank correlation coefficient, that one frequently sees 'reliability' there *defined* by the correlation coefficient between test–retest data. In the same sense, for binary data the reliability coefficient is *defined* by the intraclass kappa.

The original introductions of kappa [3, 40] defined not the population parameter, κ, but the sample estimate k, where the probability of agreement is replaced by the observed proportion of agreement, and P is estimated by the proportion of the classifications that selected category 1. This was proposed as a measure of reliability long before it was demonstrated that it satisfied the classical definition of reliability [31]. Fortunately, the results were consistent. However, that sequence of events spawned part of the problems surrounding kappa, since it opened the door for others to propose various sample statistics as measures of binary reliability, without demonstration of the relationship of their proposed measure with reliability as technically defined. Unless such a statistic estimates the same population parameter as does the intraclass kappa, it is *not* an estimate of the reliability of a binary measure. However, there are other statistics when $M = 2$, that estimate the same parameter in properly designed reliability studies (random sample from the population of subjects, and a random sample of blinded raters/ratings for each subject), such as all weighted kappas (not the same as an intraclass kappa as will be seen below), or the sample phi coefficient, the risk difference or

the attributable risk. Typically these provide less efficient estimators than does the sample intraclass kappa.

It is useful to note that $\kappa = 0$ indicates either that the heterogeneity of the patients in the population is not well detected by the raters or ratings, or that the patients in the population are homogeneous. Consequently it is well known that it is very difficult to achieve high reliability of any measure (binary or not) in a very homogeneous population (P near 0 or 1 for binary measures). That is not a flaw in kappa [26] or any other measure of reliability, or a paradox. It merely reflects the fact that it is difficult to make clear distinctions between the patients in a population in which those distinctions are very rare or fine. In such populations, 'noise' quickly overwhelms the 'signals'.

2.2. The $K \times 2$ intraclass kappa

When there are more than two categories ($K > 2$) both $\mathbf{X}_i$ and $\mathbf{p}_i$ are K-dimensional vectors. The classical definition of reliability requires that the covariance matrix of $\mathbf{p}_i$, Σ_p, be compared with the covariance matrix of $\mathbf{X}_i$, Σ_X. The diagonal elements of Σ_p are $\kappa_j P_j P_j'$, where κ_j is the 2×2 intraclass kappa with category j versus 'not-j', a pooling of the remaining categories, P_j is the $E(p_{ij})$, $P_j' = 1 - P_j$. The off-diagonal elements are $\rho_{jj^*} P_j P_{j^*}$, $j \neq j^*$, with ρ_{jj^*} the correlation coefficient between p_{ij} and p_{ij^*}. The diagonal elements of Σ_X are $P_j P_j'$, and the off-diagonal elements are $-P_j P_{j^*}$.

What has been proposed as a measure of reliability is the $K \times 2$ intraclass kappa

$$\kappa = \text{trace}(\Sigma_p)/\text{trace}(\Sigma_X) = \Sigma(P_j P_j' \kappa_j)/\Sigma(P_j P_j')$$

Again it can be demonstrated that this is equivalent to PACC with p_0 again the probability of agreement, now with $p_c = \Sigma P_j P_j'$.

From the above, it is apparent that to obtain a non-zero $K \times 2$ intraclass kappa requires that only one of the K categories have non-zero κ_j. If that one category has reasonable heterogeneity in the population ($P_j P_j'$ large) and has large enough κ_j, the $K \times 2$ intraclass kappa may be large.

Consider the special case for $K = 3$, when $\mathbf{p}_i = (1, 0, 0)$ with probability π, and $\mathbf{p}_i = (0, 0.5, 0.5)$ with probability $\pi' = 1 - \pi$. In this case category 1 is completely discriminated from categories 2 and 3, but the decisions between 2 and 3 are made randomly. Then $\kappa_1 = 1$, and $\kappa_2 = \kappa_3 = \pi/(\pi + 1)$, and the 3×2 intraclass kappa is $3\pi/(3\pi + 1)$. When $\pi = 0.5$, for example, $\kappa = 0.60$, and $\kappa_2 = \kappa_3 = 0.33$, even if 2 and 3 are here randomly assigned. Such a large overall κ can be mistakenly interpreted as a good reliability for all three categories, where here clearly only category 1 is reliably measured.

No one index, the $K \times 2$ intraclass kappa or any other, clearly indicates the reliability of a multi-category X. For categorical data, one must consider not only how distinct each category is from the pooled remaining categories (as reflected in the κ_j, $j = 1, 2, \ldots, K$), but how easily each category can be confused with each other [13, 41]. Consequently, we would suggest that: (i) multi-category kappas are not used as a measure of reliability with $K > 2$ categories; (ii) that seeking any *single* measure of multi-category reliability is a vain effort; and (iii) at least the K individual category κ_j's be reported, but that, better yet, methods be further developed to evaluate the entire misclassification matrix [42]. In particular, the decision to recommend kappa with two categories, but to recommend against kappa with more than two categories, is not influenced by the fact that kappa is related to PACC in both cases.

Table I. Estimation of the $2\times M$ intraclass correlation coefficient in the Periyakoil *et al.* data, with s the number of positive (grief) classifications from the $M=4$ raters, f_s the proportion of items with that number, k_s the kappa coefficient based on omitting one subject with s positive classifications, and w_s the weight needed to calculate the asymptotic variance.

s	f_s	s/M	$1-s/M$	k_s
0	0.2029	0.0000	1.0000	0.5700
1	0.1739	0.2500	0.7500	0.5860
2	0.0870	0.5000	0.5000	0.5918
3	0.1159	0.7500	0.2500	0.5873
4	0.4203	1.0000	0.0000	0.5725

General formula for k: $k=1-M\Sigma f_s(sM)(1-sM)/((M-1)PP')$
$P=\Sigma f_s(sM)$

Jack-knife formulae: Jack-knife $k=Nk-(N-1)\text{average}(k_s)$
Jack-knife $\text{SE}^2=(N-1)^2 s_k^2/N$
$s_k^2=\text{sample variance}(k_s)$

Results from above case: $P=0.5942$
$k=0.5792$
Jack-knife $k=0.6429$

The 2×2 intraclass kappa seems ideal as a measure of binary reliability, but the $K\times 2$ intraclass kappa we recommend against as uninterpretable. What if one had only two categories, but $M>2$ raters?

2.3. The $2\times M$ (multi-rater) intraclass kappa

With only two categories, the reliability coefficient is still $\kappa=\sigma^2/PP'$, as shown above. The multi-rater sample kappa statistic [43] is based on comparing pairwise agreement among the $M(M-1)/2$ pairs of raters evaluating each patient with what would be expected if classifications were randomly made. This process has been shown to obtain the equivalent result as applying the formula for the intraclass ρ for interval data to these binary data [44]. This statistic estimates the same κ as does the 2×2 intraclass kappa. For a fixed sample size of subjects, the larger the M, the smaller the estimation error.

There are several ways to estimate intraclass kappa here, but the easiest both for theory and application requires that the data be organized by s, the number of positive (category 1) classifications per patient (See Table I, column 1), $s=0,1,2,\ldots,M$. The proportion of the N patients sampled who have s of the M categorizations positive is f_s. The formula for calculation is presented in Table I, along with a demonstration of the calculation of this statistic based on a study conducted by one of the authors (VSP).

In this case, $N=69$ items were sampled from the population of items that might be used to distinguish preparatory grief (category 1) from depression (category 2) in dying adult patients. The issue was to assess to what extent clinicians could reliably distinguish the two. Depression, when it exists, is hypothesized to diminish quality of the dying process but can be effectively treated, while normal preparatory grief, when it exists, is hypothesized to be a sign of positive coping with the dying process that should be facilitated. $M=4$ expert clinicians were sampled and complied with classifying each item as more indicative of preparatory grief or depression. The results appear in Table I, with $k=0.579$.

Table II. The population probability distribution of the number of positive responses with M raters, generated from the sensitivity/specificity model (model A: Se = 0.60, Sp = 0.99, $\pi = 0.1525$) and the know/guess model (model B: $\pi_1 = 0.0250$, $\pi_0 = 0.7875$, $\alpha = 0.4054$). Both models have $P = 0.10$ and $\kappa = 0.50$ to two decimal places. Implication: the distribution of responses for $M > 2$ differ even when P and κ are the same.

Number of positive = s	$M = 2$		$M = 4$		$M = 6$	
	A	B	A	B	A	B
0	85.5%	85.4%	81.8%	81.1%	79.8%	79.6%
1	9.0%	9.0%	5.6%	6.4%	5.4%	3.4%
2	5.5%	5.6%	5.3%	6.5%	2.2%	5.8%
3			5.3%	3.0%	4.2%	5.2%
4			2.0%	3.0%	4.7%	2.7%
5					2.8%	0.7%
6					0.7%	2.6%

While the standard error is known and easily accessible when $M = 2$ [43, 45–47], to date when $M > 2$ it is known and easily accessible only under the null hypothesis of randomness [43]. The calculation of the standard error in general when $M > 2$ was described by Fleiss as 'too complicated for presentation' (reference [43], p. 232), referring readers to Landis and Koch [48]. Not only is this standard error difficult to access, but also it is not known exactly how accurate it is for small to moderate sample size. Part of the problem lies in attempting to obtain a general solution when there are more than two categories (where intraclass kappa may be misleading), and when the number of ratings per patient is itself a variable from patient to patient (which may be problematic). The situation with the $2 \times M$ intraclass kappa is much simpler.

For patient i, with probability p_{i1}, the probability that s of the interchangeable independent M ratings will be positive is the binomial probability $(s = 0, 1, 2, \ldots, M)$ with probability p_{i1} the binomial probability (say $\text{Bin}(s; p_{i1}, M), s = 0, 1, 2, \ldots, M$). The probability that a randomly sample subject will be positive is the expected value of $\text{Bin}(s; p_{i1}, M)$ over the unknown distribution of p_{i1}. This involves moments of the p_{i1} distribution up to order M. Since P and κ involve only the first two moments, the distribution of the number of positive responses is determined by P and κ only when $M = 2$. Consequently the quest for a standard error of the $2 \times M$ intraclass sample kappas for $M > 2$ that involves only parameters P and κ, that is, only moments up to order 2, is one of those futile quests [49]. One might have many different distributions of p_{i1} that have the same first two moments (P and κ) but that differ in the higher moments. For each such distribution the sample distribution for the $2 \times M$ intraclass sample kappa would differ. This fact differentiates the distribution theory of the intraclass kappa for binary data from that of the intraclass correlation coefficient, ρ, to which it is closely computationally related, for interchangeable normal variates, for in the latter case, the distribution is determined by ρ, however large the number of raters, M.

For example, in Table II, we present an example of a 'sensitivity/specificity' model and of a 'know/guess' model selected to have almost exactly the same $P = 0.10$ and $\kappa = 0.50$, and show the distribution of response for $M = 2, 4, 6$. It can be seen that the population distributions are almost the same for $M = 2$, slightly different for $M = 4$ and very different for $M = 6$. Thus,

unless $M=2$, one would not expect that the distributions of the $2\times M$ intraclass kappa would be the same in these two cases, much less in all cases with $P=0.10$ and $k=0.50$.

The vector of observed frequencies of the numbers of positive responses has a multinomial distribution with probabilities determined by the expected values of $\mathrm{Bin}(s{:}\,p_i, M)$. Thus one can use the methods derived by Fisher [50] to obtain an approximate (asymptotic) standard error of kappa. An approximate standard error of k can also be obtained very easily using jack-knife procedures omitting one patient at time [45, 47, 51–53], as shown in Table I. These results correspond closely to those derived in various ways for the 2×2 intraclass kappas [43, 46, 47, 54]. The jack-knife procedure is demonstrated in Table I. (As a 'rule of thumb', the minimum number of patients should exceed both $10/P$ and $10/P'$. When $P=0.5$, 20 patients are minimal; when $P=0.01$, no fewer than 1000 patients are needed.) A generalized version of the SAS program (SAS Institute Inc., Cary NC) that performs the calculations can be located at http://mirecc.stanford.edu

When there are a variable number of raters per patient, the problem becomes more complicated, since the exact distribution of responses changes as M varies, involving more or fewer unknown moments of the p_{i1} distribution. If the patient's number of ratings is totally independent of his/her p_{i1}, one could stratify the patients by the number of ratings, obtain a 2×2 intraclass kappa from those with $M=2$, a 2×3 intraclass kappa from those with $M=3$ etc., and a standard error for each. Since these are independent samples from the same parent population, one could then obtain a weighted average of the kappas and its standard error using standard methods.

However, often the variation of the number of ratings is related to p_{i1}. Patients with more serious illnesses, for example, are more likely to have a positive diagnosis and less likely to provide the greater number of ratings. In that case, the subsamples of patients with $2, 3, 4, \ldots$ ratings may represent different populations and thus have different reliabilities that should not be muddled. This raises some serious questions about the practical application of the standard error derived by Landis and Koch [48] or any solution in which the number of ratings is variable.

To summarize, for the purpose of measuring reliability of a binary measure, the $2\times M$ ($M\geqslant 2$) is highly recommended, but the use of the $K\times M$ kappa for $K>2$ is questionable. To this it should be added that useful standards have been suggested for evaluation of the $2\times M$ kappa as a measure of reliability [24], with $k\leqslant 0.2$ considered slight, $0.2<k\leqslant 0.4$ as fair; $0.4<k\leqslant 0.6$ as moderate, $0.6<k\leqslant 0.8$ as substantial and $k>0.8$ as almost perfect. It is important to realize that a kappa coefficient below 0.2 is slight, no matter what the p-value is of a test of the null hypothesis of randomness. Moreover, a kappa coefficient above 0.6 that is not 'statistically significant' on such a test indicates inadequate sample size, not a definitive conclusion about the reliability of the measure. It is the magnitude of k that matters, and how precisely that is estimated, not the p-value of a test of the null hypothesis of randomness [55].

3. VALIDITY OF CATEGORICAL MEASURES: THE $K\times M$ WEIGHTED KAPPAS

The validity of a measure is defined as the proportion of the observed variance that reflects variance in the construct the measure was intended to measure [36, 38], and is thus always no greater than the reliability of a measure. Validity is generally assessed by a correlation

coefficient between a criterion or 'gold standard' (X_i) and the measure (Y_i) for each patient in a representative sample from the population to which the results are to be generalized. (Once again, a measure might be more valid in one population than in another.) If a measure is completely valid against a criterion, there should be a 1:1 mapping of the values of Y_i onto the values of X_i. With categorical measures, the hope is to be able to base clinical or research decisions on Y_i that would be the same as if those decisions were based on the 'gold standard' X_i. That would require not only that the number of categories of Y_i match the number of categories of X_i, but that the labels be the same.

The 'gold standard' is the major source of difficulty in assessing validity, for there are very few true 'gold standards' available. Instead, many 'more-or-less gold standards' are considered, each somewhat flawed, but each of which provides some degree of challenge to the validity of the measure. Thus, as in the case of reliability, there are many types of validity, depending on how the 'gold standard' is selected: face validity; convergent validity; discriminative validity; predictive validity; construct validity.

While there are many problems in medical research that follow this paradigm, few of which are actually labelled 'validity' studies, we will for the moment focus on medical test evaluation. In medical test evaluation, one has a 'gold standard' evaluation of the presence/absence or type of disease, usually the best possible determination currently in existence, against which a test is assessed. To be of clinical and policy importance the test result for each patient should correspond closely to the results of the 'gold standard', for treatment decisions for patients are to be based on that result.

3.1. A 2 × 2 weighted kappa coefficient

Once again the most common situation is with two ratings per patient, say X_i and Y_i each having only two categories of response. We use different designations here for the two ratings, X_i and Y_i, in order to emphasize that the decision process underlying the 'gold standard' (X_i) and the diagnosis under evaluation (Y_i) are, by definition, not the same. For the same reason, we focus on the probability of a positive result (category 1) in each case, with probability p_{i1} for X_i and q_{i1} for Y_i, using different notation for the probabilities.

The distribution of p_{i1} and q_{i1} in the population of patients may be totally different, even if $P = E(p_{i1})$ and $Q = E(q_{i1})$ are equal. The equality of P and Q cannot be used to justify the use of the intraclass kappa in this situation, for the intraclass kappa is appropriate only to the situation in which all the moments, not just the first, are equal (interchangeable variables).

Since X_i and Y_i are 'blinded' to each other, the probability that for patient i both X_i and Y_i are positive is $p_{i1}q_{i1}$. Thus in the population, the probability that a randomly selected patient has both X_i and Y_i positive is $E(p_{i1}q_{i1}) = PQ + \rho\sigma_p\sigma_q$, where $P = E(p_{i1})$, $Q = E(q_{i1})$, ρ is the product moment correlation coefficient between p_{i1} and q_{i1}, $\sigma_p^2 = \text{variance}(p_{i1})$, $\sigma_q^2 = \text{variance}(q_{i1})$. All the probabilities similarly computed are presented in Table III.

It can be seen in Table III that the association between X_i and Y_i becomes stronger as $\rho\sigma_p\sigma_q$ increases from zero. At zero, the results in the table are consistent with random decision making. Any function of $\rho\sigma_p\sigma_q$, P and Q, that is strictly monotonic in $\rho\sigma_p\sigma_q$, that takes on the value zero when $\rho = 0$, and takes on the value $+1$ when the probabilities on the cross diagonal are both 0, and -1 when the probabilities on the main diagonal are both 0, is a type of correlation coefficient between X and Y. The difficulty is that there are an infinite number of such functions (some of the most common defined in Table III), and therefore an

Table III. The 2×2 weighted kappa: probabilities and weights. Definitions of some common measures used in medical test evaluation or in risk assessment.

	$Y=1$	$Y=2$	Total
Probabilities			
$X=1$	$a=PQ+\rho\sigma_p\sigma_q$	$b=PQ'-\rho\sigma_p\sigma_q$	P
$X=2$	$c=P'Q-\rho\sigma_p\sigma_q$	$d=P'Q'+\rho\sigma_p\sigma_q$	$P'=1-P$
Total	Q	$Q'=1-Q$	
Weights indicating loss or regret $(0<r<1)$:			
$X=1$	0	r	
$X=2$	$r'=1-r$	0	

$\kappa(r)=(ad-bc)/(PQ'r+P'Qr')=\rho\sigma_p\sigma_q/(PQ'r+P'Qr'),(0<r<1)$.
$\kappa(1/2)=2(ad-bc)/(PQ'+P'Q)=(p_0-p_c)/(1-p_c),(p_0=a+d,\ p_c=PQ+P'Q')$.

Sensitivity of Y to X: $\mathrm{Se}=a/P=Q+Q'\kappa(1)$.
Specificity of Y to X: $\mathrm{Sp}=d/P'=Q'+Q\kappa(0)$.
Predictive value of a positive test: $\mathrm{PVP}=a/Q=P+P'\kappa(0)$.
Predictive value of a negative test: $\mathrm{PVN}=d/Q'=P'+P\kappa(1)$.
Percent agreement $=p_0=a+d=p_c+p_c'\kappa(1/2)$.
Risk difference $=\mathrm{Se}+\mathrm{Sp}-1=a/P-c/P'=\kappa(Q')$.
Attributable risk $=\kappa(0)$.
Odds ratio $=ad/bc=(\mathrm{SeSp})/(\mathrm{Se'Sp'})=(\mathrm{PVP\ PVN})/(\mathrm{PVP'\ PVN'})$.

infinite number of correlation coefficients that yield results not necessarily concordant with each other.

There is one such correlation coefficient, a certain 2×2 weighted kappa, unique because it is based on an acknowledgement that the *clinical* consequences of a false negative (X_i positive, Y_i negative) may be quite different from the *clinical* consequences of a false positive (X_i negative, Y_i positive) [47]. For example, a false negative medical test might delay or prevent a patient from obtaining needed treatment in timely fashion. If the test were to fail to detect the common cold, that might not matter a great deal, but if the test were to fail to detect a rapidly progressing cancer, that might be fatal. Similarly a false positive medical test may result in unnecessary treatment for the patient. If the treatment involved taking two aspirin and calling in the morning, that might not matter a great deal, but if it involved radiation, chemotherapy or surgical treatment, that might cause severe stress, pain, costs and possible iatragenic damage, even death, to the patient. The balance between the two types of errors shifts depending on the population, the disorder and the medical sequelae of a positive and negative test. This weighted kappa coefficient is unique among the many 2×2 correlation coefficients in that in each context of its use, it requires that this balance be explicitly assessed *a priori* and incorporated into the parameter.

For this particular weighted kappa, a weight indicating the clinical cost of each error is attributed to each outcome (see Table III); an index r is set that ranges from 0 to 1 indicating the relative importance of false negatives to false positives. When $r=1$, one is primarily concerned with false negatives (as with a screening test); when $r=0$, one is primarily concerned with false positives (as with a definitive test); when $r=1/2$, one is equally concerned with both (as with a discrimination test). The definition of $\kappa(r)$ in this case [47, 56] is

$$\kappa(r)=\rho\sigma_p\sigma_q/(PQ'r+P'Qr')$$

The sample estimator is $k(r) = (ad - bc)/(PQ'r + P'Qr')$, where a, b, c, d are the proportions of the sample in the cells so marked in Table III, P and Q estimated by the sample proportions. Cohen's kappa [40], often called the 'unweighted' kappa, is $\kappa(1/2)$

$$\kappa(1/2) = (p_0 - p_c)/(1 - p_c)$$

where p_0 again is the proportion of agreement, and here $p_c = PQ + P'Q'$, once again a PACC (see Table III for a summary of definitions). When papers or programs refer to 'the' kappa coefficient, they are almost inevitably referring to $\kappa(1/2)$, but it must be recognized that $\kappa(1/2)$ reflects a decision (conscious or unconscious) that false negatives and false positives are equally clinically undesirable, and $\kappa(r)$ equals PACC only when $r = 1/2$.

Different researchers are familiar with different measures of 2×2 association, and not all readers will be familiar with all the following. However, it is important to note the strong interrelationships among the many measures of 2×2 association. Risk difference (Youden's index) is $\kappa(Q')$, and attributable risk is $\kappa(0)$, reflecting quite different decisions about the relative importance of false positives and negatives. The phi coefficient is the geometric mean of $\kappa(0)$ and $\kappa(1)$: $(\kappa(0)\kappa(1))^{1/2}$. Sensitivity and predictive value of a negative test rescaled to equal 0 for random decision making and 1 when there are no errors, equal $\kappa(1)$. The specificity and predictive values of a positive test, similarly rescaled, equal $\kappa(0)$. For any r between 0 and 1, $\kappa(r)/\max \kappa(r)$ and phi/max phi [57], where max $\kappa(r)$ and max phi are the maximal achievable values of $\kappa(r)$ and phi, respectively, equal either $\kappa(0)$ or $\kappa(1)$, depending on whether P is greater or less than Q. This briefly demonstrates that most of the common measures of 2×2 association either (i) equal $\kappa(r)$ for some value of r, or, (ii) when rescaled, equal $\kappa(r)$ for some value of r, or (iii) equal some combination of the $\kappa(r)$. Odds ratio and measures of association closely related to odds ratio seem the notable exceptions.

Researchers sometimes see the necessity of deciding *a priori* on the relative clinical importance of false negatives versus false positives as a problem with $\kappa(r)$, since other measures of 2×2 association do not seem to require any such *a priori* declaration. In fact, the opposite is true. It has been demonstrated [58] that every measure of 2×2 association has implicit in its definition some weighting of the relative importance of false positives and false negatives, often unknown to the user. The unique value of this weighted kappa as a measure of validity is that it *explicitly* incorporates the relative importance of false positives and false negatives, whereas users of other 2×2 measures of association make that same choice by choosing one measure rather than another, and often do so unaware as to the choice they have *de facto* made. If they are unaware of the choice, that is indeed a problem, for there is risk of misleading clinical and policy decisions in the context in which the user applies it [58].

However, unlike the situation with reliability, it cannot be argued that $\kappa(r)$, in any sense, defines validity, for the appropriate choice of a validity measure depends on what the user stipulates as the relative importance of false positives and false negatives. How these are weighted may indicate a choice of index not directly related to any $\kappa(r)$ (the odds ratio, for example).

It is of importance to note how the relative clinical importance (r) and the reliabilities of X and Y (the intraclass κ_X and κ_Y defined above for X and Y) influence the magnitude of $\kappa(r)$:

$$\kappa(r) = \rho(\kappa_X \kappa_Y)^{1/2}(PP'QQ')^{1/2}/(PQ'r + P'Qr')$$

with $P' = 1 - P,\ Q' = 1 - Q,\ r' = 1 - r$.

Here, as defined above, ρ is the correlation between p_{i1} and q_{i1} (which does not change with r). κ_X and κ_Y are the test–retest reliabilities of X and Y (which do not depend on r). As is always expected of a properly defined reliability coefficient, the correlation between X and Y reflected in $\kappa(r)$ suffers attenuation due to the unreliabilities of X and Y, here measured by the intraclass kappas κ_X and κ_Y. Only the relationship between P and Q affects $\kappa(r)$ differently for different values of r. When $P=Q$, $\kappa(r)$ is the same for all values of r and estimates the same population parameter as does the intraclass kappa although the distribution of the sample intraclass kappa is not exactly the same as that of the sample weighted kappa. For that matter, when $P=Q$, the sample distributions of $k(r)$ for different values of r are not all the same, even though all estimate the same parameter. Otherwise, in effect, too many positive tests ($Q>P$) are penalized by $\kappa(r)$ when false positives are of more concern (r nearer 0), and too many negative tests ($Q<P$) are penalized by $\kappa(r)$ when false negatives are of more concern (r nearer 1).

A major source of confusion in the statistical literature related to kappa is the assignment of weights [13]. Here we have chosen to use weights that indicate loss or regret, with zero loss for agreements. Fleiss [43] used weights that indicate gain or benefit, with maximal weights of 1 for agreements. Here we propose that false positives and false negatives may have different weights. Fleiss required that they be the same. Both approaches are viable for different medical research problems, as indeed are many other sets of weights, including sets that assign different weights to the two types of agreements.

If the weights reflect losses or regrets, $\kappa(r)=(E_c(r)-E_o(r))/(E_c(r)-\min)$, while if the weights reflect gains or benefits, $\kappa(r)=(E_o(r)-E_c(r))/(\max-E_c(r))$, where $E_c(r)$ is the expected weight when $\rho=0$ and $E_o(r)$ the expected weight with the observed probabilities. The scaling factor min is the ideal minimal value of $E_o(r)$ when losses are considered, and max is the ideal maximal value of $E_o(r)$ when gains are considered, for the particular research question. Here min is 0, where there are no disagreements; Fleiss' max is 1, also when there are no disagreements. Regardless of the weight assigned to disagreements in Fleiss' version of kappa, his weighted kappas in the 2×2 situation all correspond to what is here defined as $\kappa(1/2)$, while if P and Q are unequal, here $\kappa(r)$ changes with r, and generally equals $\kappa(1/2)$ only when $r=1/2$.

How the weights, min and max, are assigned changes the sampling distribution of $\kappa(r)$, which may be one of the reasons finding its correct standard error has been so problematic. Since the weights should be dictated by the nature of the medical research question, they should and will change from one situation to another. It is not possible to present a formula for the standard error that would be correct for all possible future formulations of the weights. For the particular weights used here (Table III) the Fisher procedure [50] could be used to obtain an asymptotic standard error. However, given the difficulties engendered by the wide choice of weights, and the fact that it is both easier and apparently about as accurate [54] when sample size is adequate, we would here recommend instead that the jack-knife estimator be used. That would guarantee that the estimate of the standard error be accurate for the specific set of weights selected and avoid further errors.

3.2. The $K\times 2$ multi-category kappa

In the validity context, as noted above, if the 'gold standard' has K categories, any candidate valid measure must also have K categories with the same labels. Thus, for example,

Table IV. Example: the joint probability distribution of a three-category X and a three-category Y, with one perfectly valid category ($Y=1$ for $X=1$), and two invalid categories ($Y=2$ for $X=2$) and ($Y=3$ for $X=3$) because of an interchange of $Y=2$ and $Y=3$ ($P_1+P_2+P_3=1$).

	$Y=1$	$Y=2$	$Y=3$	Total
$X=1$	P_1	0	0	P_1
$X=2$	0	0	P_2	P_2
$X=3$	0	P_3	0	P_3
Total	P_1	P_3	P_2	

if the 'gold standard' identifies patients with schizophrenia, depression, personality disorder, and 'other', any potentially valid diagnostic test would also identify the same four categories. In a direct generalization of the above, if 'gold standard' and diagnosis agree, disagreement is zero. If, however, someone who is schizophrenic is treated for depression, that is not an error necessarily of equal clinical importance as someone who is depressed being treated for schizophrenia. For each possible disgreement, one could assess the relative clinical importance of that misclassification, denoted r_{jj^*} for $j \neq j^*$. The only requirement is that $r_{jj^*} \geqslant 0$ for all $j \neq j^*$, and that $\Sigma r_{jj^*} = 1$. Then the weighted kappa, $\kappa(r)$, is defined as above as $(E_c(r) - E_o(r))/E_c(r)$.

The difficulty here, as with the $K \times 2$ intraclass kappa, is that $\kappa(r)$ is sure to equal 0 only if *all* classifications are random. Thus having only one valid category can yield a positive $\kappa(r)$, or we might have $\kappa(r)$ near zero when all but one category are completely valid.

For example, consider the case shown in Table IV. Here diagnostic category 1 is completely valid for 'gold standard' category 1, but diagnostic categories 2 and 3 are obviously switched. When (all r_{jj^*} here equal) $P_1 = 0.1$, $P_2 = 0.4$ and $P_3 = 0.5$, $k(r) = -0.525$. When $P_1 = 0.3$, $P_2 = 0.5$, $P_3 = 0.2$, $k(r) = +0.014$. When $P_1 = 0.8$, $P_2 = P_3 = 0.1$, $k(r) = +0.412$. None of these results $(-0.525, +0.014, +0.412)$ suggests what is obvious from examination of the complete cross-classification matrix: Y-categories 2 and 3 must be switched to obtain perfect validity. Consequently, once again, we propose that, like the multi-category intraclass kappa, the multi-category weighted kappas not be used as a measure of validity, for no single measure of validity can convey completely and accurately the validity of a multi-category system, where some categories may be valid but vary in terms of degree of validity, and others may be invalid.

3.3. *The $2 \times M$ Multi-rater kappa*

Now suppose that we had a binary 'gold standard' X_i, and M binary diagnostic tests: $Y_{i1}, Y_{i2}, \ldots, Y_{iM}$. Can the M diagnostic tests be used to obtain a valid diagnosis of X_i, and how valid would that test be? In this case, X_i and each Y_{ij} may have a different underlying distribution of p_{i1} or q_{i1}. While we could propose a multi-rater kappa [59], generally the way this problem is approached in medical test evaluation is by developing a function $g(Y_{i1}, Y_{i2}, \ldots)$, called a 'risk score', such that $g()$ is monotonically related to $\text{Prob}(X_i = 1)$. Then some cutpoint is selected so that if $g(Y_{i1}, Y_{i2}, \ldots) \geqslant C$, the diagnostic test is positive, and otherwise negative.

Almost inevitably, applying such a cutpoint dichotomizing the ordinal risk score to a binary classification reduces the power of statistical tests based on the measures [60]. If the cutpoint is

injudiciously chosen, it may also mislead research conclusions. However, for clinical decision making, that is, deciding who to treat and not treat for a condition, who to hospitalize or not, a binary measure is necessary. Thus while the recommendation not to dichotomize for purposes of research is almost universal, dichotomization for clinical purposes is often necessary. Such dichotomization reduces the multivariate tests to a binary test based on all the individual tests. The 2×2 weighted kappa may then be used as a measure of the validity of the combined test.

The most common method of developing this function is multiple logistic regression analysis where it is assumed that logit $\text{Prob}(X_i = 1 | Y_{i1}, Y_{i2}, \ldots) = \beta_0 + \Sigma \beta_j Y_{ij}$, that is, some linear function of the Y's, with a 'risk score' ($\Sigma \beta_j Y_{ij}$) assigned to each patient. Regression trees [56, 61] can also be used, using whatever validity criterion the developer chooses to determine the optimal test at each stage and a variety of stopping rules. Each patient in a final branch is given a 'risk score' equal to the $\text{Prob}(X_i = 1)$ in that subgroup. Finally, one might simply count the number of positive tests for each patient, $g(Y) = \Sigma Y_{ij}$, and use this as a 'risk score'. There are many such approaches, all of which reduce the 2^M possible different responses to the M binary tests to a single ordinal response, the 'risk score', using all M tests in some sense optimally. The relative strengths and weaknesses of these and other approaches to developing the 'risk score' can be vigorously debated. However, that is not the issue here.

When the 'risk score' is determined, the cutpoint C is often selected to equate P and Q, that is, so that $Q = \text{Prob}(g(Y_{i1}, Y_{i2}, \ldots) \geqslant C) = P$. This is not always ideal. Better yet, the optimal cutpoint would be the one that maximizes $\kappa(r)$, where r again indicates the relative importance of false negatives to false positives [56], or whichever other measure of 2×2 association best reflects the trade-offs between false positives and false negatives.

We do not recommend any $2 \times M$ weighted kappa coefficient as a measure of validity, for there are already a variety of other standard methods used in this problem that seem to deal well with the problem. None seems to require or would benefit from a $2 \times M$ kappa coefficient, for all focus more appropriately on reducing the problem to a 2×2 problem. Then the 2×2 weighted kappa might be used as a measure of validity.

4. THE PROBLEM OF CONSENSUS DIAGNOSIS

The final context of medical research in which kappa coefficients have proved uniquely useful is that of the consensus diagnosis. Suppose one assesses the reliability of a binary X_i, and found that its reliability, as measured by a $2 \times M$ intraclass kappa, was greater than zero, but not satisfactory. 'Rule of thumb' standards for reliability have been proposed [14, 24]. By those standards, $\kappa = 0.579$, as in the Periyakoil data, or $\kappa = 0.5$, as in both cases of Table II, would be considered 'moderate' [24] or 'fair' [14]. Could one use a consensus of M raters, requiring at least C positive diagnoses for a consensus positive diagnosis, and thereby achieve adequate (say $\kappa > 0.8$, 'almost perfect' or 'substantial') reliability? How large should M be, and what value of C should be chosen?

One could deal with the problem using brute force: sample $2M$ raters for each patient sampled, randomly split the raters into two groups of M for each patient. Then for $C = 1, 2, \ldots, M$, determine the diagnosis for that value of C, and obtain 2×2 intraclass kappa, κ_{CM}. Then choose the optimal cutpoint C as the one that maximizes κ_{CM} for that value of M. Then vary M.

Table V. The optimal consensus diagnoses for the sensitivity/specificity model with $Se = 0.60$. $Sp' = 0.01$, $\pi = 0.1525$, and for the know/guess model with $\pi_1 = 0.0250$, $\pi_0 = 0.7875$, $\alpha = 0.4054$. Both models have $P = 0.10$, $\kappa = 0.50$. The number of diagnoses in the consensus is M, with C the optimal cutpoint (a positive diagnosis is given those with C or more positive diagnoses of the M). Q is the proportion diagnosed positive with the optimal consensus, and κ is the intraclass κ for that consensus.

M	Sensitivity/specificity model			Know/guess model		
	C	Q	κ	C	Q	κ
1	1	0.10	0.50	1	0.10	0.50
2	1	0.14	0.70	1	0.15	0.66
3	1	0.17	0.76	1	0.17	0.78
4	2	0.13	0.79	1	0.19	0.87
5	2	0.14	0.89	1	0.20	0.92
6	2	0.15	0.94	1	0.20	0.95
7	2	0.15	0.96	1	0.21	0.97
8	2	0.15	0.97	1	0.21	0.98
9	2	0.15	0.97	1	0.21	0.99
10	3	0.15	0.98	1	0.21	0.99

With the four raters in Table I, we have already calculated that $\kappa_{11} = 0.579$. We then randomly split the pool of four raters into two sets of two for each patient, and found that $\kappa_{12} = 0.549$, and $\kappa_{22} = 0.739$. Thus the optimal consensus of 2 is to use a cutpoint $C = 2$, and the reliability then rises from $\kappa_{11} = 0.579$ with one rater to $\kappa_{22} = 0.739$ for an optimal consensus of two. For an expanded discussion of these methods, see Noda *et al.* [62], and for a program to perform such calculations see http://mirecc.stanford.edu

It is of note that if the optimal consensus of 2 is obtained when $C = 1$, in practice one would not request a second opinion when the first one was positive. If, as above, the optimal consensus of 2 is obtained when $C = 2$, in practice one would not request a second opinion when the first one was negative. It often happens with the optimal consensus that, when put into practice, the number of ratings per patient to obtain a consensus of M is far less than M ratings per patient. This often means one can increase the quality of the diagnosis with minimal increase in time and cost. However, to identify that optimal consensus in the first place requires $2M$ ratings for each patient. Thus to evaluate a consensus of 3, one needs 6 ratings per patient, for 4, one needs 8, etc. This rapidly becomes an unfeasible solution in practice.

The theoretical solution is easy. For a patient with probability p_{i1} on a single rating, the probability of a positive diagnosis for a consensus of C of M is

$$q_{iCM} = \mathrm{Bin}(C; p_{i1}, M)$$

where $\mathrm{Bin}(C; p_{i1}, M)$ is the probability that a binomial random variable with parameters p_i and M equals or exceeds C. Thus $Q_{CM} = E(\mathrm{Bin}(C; p_{i1}, M))$ and $\kappa_{CM} = \mathrm{var}(q_{iCM})/(Q_{CM}Q'_{CM})$. If we knew the distribution of p_{i1}, we would also know the distribution of q_{iCM} for all C and M, and thus know κ_{CM}. In Table V, for example, are presented the two hypothetical cases of Table II, where we do know the distribution and they have almost identical P and κ. Here for the sensitivity/specificity model, as M increases from 1 to 10, the optimal C rises

from 1 to 2 to 3, and the κ from 0.50 for one observation to 0.98 for a consensus of 10. One would need a consensus of 2 positive out of 5 to achieve $\kappa \geqslant 0.8$. On the other hand, for the 'know/guess' model, as M increases from 1 to 10, the optimal C is always equal to 1, but the κ still rises from 0.50 for one observation to 0.99 for a consensus of 10. One would now need a consensus of 1 positive out of 4 to achieve $\kappa \geqslant 0.8$.

The above illustration demonstrates the fallacy of certain intuitive notions:

(i) It is not necessarily true that the optimal consensus equates Q with P.
(ii) The 'majority rule' (always use the first C exceeding $M/2$), is not always best.
(iii) The 'unanimity rule' (always use $C=0$ or $C=M$), too, is not always best.
(iv) Knowing P and κ does not settle the issue, for quite different optimal consensus rules were derived for the two situations in Table II having almost the same P and κ.

Since the reason for dichotomization is most compelling for purposes of clinical decision making, these false intuitive notions can mislead such decisions.

Examination of cases such as these provides some insight into the solution. For the 'sensitivity/specificity' model, it can be seen that for every M, the optimal C cuts off as close to the top 15 per cent of the number of positives as is possible. That 15 per cent corresponds to the 'high risk' subgroup with $p_{i1} = \mathrm{Se} = 0.60$. For the 'know/guess' model, the optimal C cuts off as close to the top 21 per cent of the number of positives as is possible. That 21 per cent corresponds to the 'high risk' comprising the subgroup of 2.5 per cent with $p_{i1} = 1$ plus the subgroup of 18.8 per cent with $p_{i1} = 0.4054$. However, in general, what proportion Q^* constitutes the 'high risk' subgroup?

The numerator of κ is $\mathrm{var}(p_{i1})$ which, for any P^* between 0 and 1, can be partitioned into two components:

$$\mathrm{var}(p_{i1}) = 2Q^*Q^{*\prime}(\mu_1 - \mu_2)^2 + Q^*\mathrm{var}(p_{i1} | p_{i1} \geqslant P^*) + Q^{*\prime}\mathrm{var}(p_{i1} | p_{i1} < P^*)$$

where $Q^* = \mathrm{prob}(p_{i1} \geqslant P^*)$, $Q^{*\prime} = 1 - Q^*$, $\mu_1 = E(p_{i1} | p_{i1} \geqslant P^*)$, and $\mu_2 = E(p_{i1} | p_{i1} < P^*)$. The percentage cut off by optimal C approximates Q^*, for that value of P^* for which the first term of $\mathrm{var}(p_{i1})$ is maximized. Thus the optimal cutpoint for p_{i1} (P^*), which determines the percentage of 'high risk' subjects (Q^*), is determined by what dichotomization of the p_{i1} distribution absorbs as much of the variance as possible [63].

5. CONCLUSIONS

To summarize:

(i) The $2 \times M$ intraclass kappa ($M \geqslant 2$) for a well-designed reliability study directly estimates reliability as defined in the classical sense and is thus the ideal reliability coefficient for a binary measure. For reasonable sample size, its standard error can be easily computed, and used to formulate confidence intervals, to test homogeneity of κ's and to address other such statistical challenges, such as developing optimal consensus rules.
(ii) The 2×2 weighted kappa $\kappa(r)$ described here is an excellent choice as a validity measure, although not a unique choice. However, since it explicitly requires that the relative importance of false positives and false negatives be specified and incorporated

into the validity measure, while all other 2×2 measures require that choice implicitly, $\kappa(r)$ is highly recommended in this context. For reasonable sample size, its standard error can easily be computed using jack-knife methods.

(iii) The $K \times M$ intraclass kappa, for $K > 2$, is not recommended as a measure of reliability, for no single measure is sufficient to completely and accurately convey information on reliability when there are more than two categories.

(iv) The $K \times M$ weighted kappas for $K > 2$ or $M > 2$ are not recommended as validity measures. When $K > 2$, the situation is similar to that with the $K \times M$ intraclass kappa. Any single measure, including $\kappa(r)$, is not enough to provide the necessary information on validity when some categories may be valid and others not. When $M > 2$, all the preferred methods in one way or another dichotomize the multi-dimensional Y space to create a binary outcome, and may then choose to use the 2×2 weighted kappa as a measure of validity. A $K \times M$ weighted kappa is not needed.

Even limited to these two contexts of reliability and validity, a broad spectrum of important medical research problems are encompassed. The $2 \times M$ intraclass kappa applies to any situation in which units are sampled from some population, and multiple subunits are sampled from each unit, where the intra-unit concordance or the inter-unit heterogeneity is of research interest.

For example, the intraclass kappa is useful as a measure of twin concordance in genetic studies of twins [64] (and could be used for triplets or quadruplets), as a measure of inter-sibling concordance in family studies, of intra-group concordance among patients in a therapy group etc. A research question such as the following also falls into the same category: If one sampled physicians or hospitals who performed a certain procedure, and assessed the outcome (success/failure) on a random sample of M of each physician's or hospital's patients undergoing that procedure, how heterogeneous would the physicians or hospitals prove to be? Here $\kappa = 0$ would indicate absolute homogeneity of results; larger κ would indicate greater heterogeneity (perhaps related to the type of patients referred, training, skill, resources or experience). Moreover, if there were a hypothesized source of heterogeneity (perhaps those that specialize in that procedure versus those that only occasionally do it), one could stratify the population by that source, compute the $2 \times M$ intraclass kappa within each stratum. If indeed that source accounted for most of the heterogeneity, the $2 \times M$ intraclass kappa within each stratum would approach zero.

The 2×2 weighted kappa in general could be applied to any situation in which the correlation between binary X_i and binary Y_i is of interest, where there are clinical consequences to be associated with the decisions. The particular weighted kappa discussed here is particularly relevant when Y_i is to be used to make decisions relative to X_i, in which case it is prudent to consider the relative clinical importance of false positives and false negatives. There are a vast number of research questions of this type in medical research. We have used as an example the evaluation of a medical test against a binary 'gold standard'. Since such medical tests are often the basis of medical decisions of whom to treat and how, such problems are of crucial importance. However, X_i might also represent the presence or absence of a disorder, and Y_i a possible risk factor for that disorder. Such information often influences policy recommendations as to preventive measures or targeting of certain populations for preventive interventions. In that situation, $\kappa(r)$ would be used as a measure of potency of that risk factor [58]. X_i might be the diagnosis by an acknowledged expert, and Y_i the diagnosis by

a less expert clinician, a nurse, a layman, from a different source, or under different conditions. Such questions often arise in health services research, for if one can achieve the same (or better) quality of diagnosis from less costly sources, one could decrease medical costs with no decrease in quality of care. What characterizes all these situations is that X_i is a criterion against which Y_i is to be evaluated, and that there are costs to misclassification that are embodied in the weight, r, that defines $\kappa(r)$.

There are many other medical research questions for which some form of kappa could conceivably be used, but to date, the logic of suggesting any form of kappa is either absent or weak. For example, to show that two disorders are non-randomly comorbid in a population, one would assess how frequently they co-occur in that population and show this is more frequent than random association would suggest [65]. One could certainly use a kappa to measure such comorbidity, but which kappa, and why any kappa would be preferable to the odds ratio, for example, is not clear. If one were interested in whether one could use a single nominal observation plus other information to predict a second nominal observation, one might prefer various regression modelling approaches, such as log-linear models [5, 20, 66]. So far, there appear to be few other contexts not covered above, where use of a kappa coefficient might be unequivocally recommended or preferred to other methods. Thus it appears that there are certain situations where kappa coefficients are ideally suited to address research questions ($2 \times M$ intraclass kappa for reliability), certain situations in which kappa coefficients have qualities that make them outstanding choices (2×2 weighted kappa in the validity context), and many other situations in which kappa coefficients may mislead or where other approaches might be preferable.

ACKNOWLEDGEMENTS

This work was supported in part by the National Institute of Mental Health grant MH40041, the National Institute of Aging grant AG17824, the Department of Veterans Affairs Sierra-Pacific MIRECC, and the Medical Research Service of the Department of Veterans Affairs.

REFERENCES

1. Fleiss JL, Nee JCM, Landis JR. Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin* 1979; **86**:974–977.
2. Scott WA. Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly* 1955; 321–325.
3. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960; **20**:37–46.
4. Feinstein AR. A bibliography of publications on observer variability. *Journal of Chronic Diseases* 1985; **38**: 619–632.
5. Banerjee M, Capozzoli M, McSweeney L, Sinha D. Beyond kappa: a review of interrater agreement measures. *Canadian Journal of Statistics* 1999; **27**:3–23.
6. Bartko JJ, Carpenter WT. On the methods and theory of reliability. *Journal of Nervous and Mental Disease* 1976; **163**:307–317.
7. Brennan RL, Prediger DJ. Coefficient kappa; some uses, misuses, and alternatives. *Educational and Psychological Measurement* 1981; **41**:687–699.
8. Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics* 1975; **31**:651–659.
9. Green SB. A comparison of three indexes of agreement between observers: proportion of agreement, G-index, and kappa. *Educational and Psychological Measurement* 1981; **41**:1069–1072.
10. Kraemer HC, Bloch DA. Kappa coefficients in epidemiology: an appraisal of a reappaisal. *Journal of Clinical Epidemiology* 1988; **41**:959–968.

11. Landis JR, Koch GG. A review of statistical methods in the analysis of data arising from observer reliability studies (Part I). *Statistica Neerlandica* 1975; **29**:101–123.
12. Light RJ. Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological Bulletin* 1971; **76**:365–377.
13. Maclure M, Willett WC. Misinterpretation and misuse of the Kappa statistic. *American Journal of Epidemiology* 1987; **126**:161–169.
14. Shrout PE. Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research* 1998; **7**:301–317.
15. Thompson WD, Walter SD. A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology* 1988; **41**:949–958.
16. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 1993; **46**:423–429.
17. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 1990; **43**:543–549.
18. Hoehler FK. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology* 2000; **53**:499–503.
19. Lantz CA, Nebenzahl E. Behavior and interpretation of the k statistic: resolution of the two paradoxes. *Journal of Clinical Epidemiology* 1996; **49**:431–434.
20. May SM. Modelling observer agreement–an alternative to kappa. *Journal of Clinical Epidemiology* 1994; **47**:1315–1324.
21. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 1971; **76**:378–382.
22. Kraemer HC. Extensions of the kappa coefficient. *Biometrics* 1980; **36**:207–216.
23. Kupper LL. On assessing interrater agreement for multiple attribute responses. *Biometrics* 1989; **45**:957–967.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–174.
25. Janes CL. Agreement measurement and the judgment process. *Journal of Nervous and Mental Disease* 1979; **167**:343–347.
26. Spitznagel EL, Helzer JE. A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry* 1985; **42**:725–728.
27. Cicchetti DV, Heavens RJ. A computer program for determining the significance of the difference betweeen pairs of independently derived values of kappa or weighted kappa. *Educational and Psychological Measurement* 1981; **41**:189–193.
28. Donner A, Eliasziw M. Sample size requirements for reliability studies. *Statistics in Medicine* 1987; **6**: 441–448.
29. Donner A, Eliasziw M, Klar N. Testing the homogeneity of kappa statistics. *Biometrics* 1996; **52**:176–183.
30. Donner A. Sample size requirements for the comparison of two or more coefficients of inter-observer agreement. *Statistics in Medicine* 1998; **17**:1157–1168.
31. Kraemer HC. Ramifications of a population model for k as a coefficient of reliability. *Psychometrika* 1979; **44**:461–472.
32. Aickin M. Maximum likelihood estimation of agreeement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics* 1990; **46**:293–302.
33. Maxwell AE. Coefficients of agreement between observers and their interpretations. *British Journal of Psychiatry* 1977; **130**:79–83.
34. Kraemer HC. Estimating false alarms and missed events from interobserver agreement: comment on Kaye. *Psychological Bulletin* 1982; **92**:749–754.
35. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam J. *The Dependability of Behavioral Measurements*. Wiley: New York, 1972.
36. Lord FM, Novick MR. *Statistical Theories of Mental Test Scores*. Addison-Wesley: Reading, MA, 1968.
37. Kraemer HC. Measurement of reliability for categorical data in medical research. *Statistical Methods in Medical Research* 1992; **1**:183–199.
38. Carey G, Gottesman II. Reliability and validity in binary ratings. *Archives of General Psychiatry* 1978; **35**: 1454–1459.
39. Huynh H. Reliability of multiple classifications. *Psychometrika* 1978; **43**:317–325.
40. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 1968; **70**:213–229.
41. Darroch JN, McCloud PI. Category distinguishability and observer agreement. *Australian Journal of Statistics* 1986; **28**:371–388.
42. Donner A, Eliasziw MA. A hierarchical approach to inferences concerning interobserver agreement for multinomial data. *Statistics in Medicine* 1997; **16**:1097–1106.
43. Fleiss JL. *Statistical Methods For Rates and Proportions*. Wiley: New York, 1981.
44. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 1973; **33**:613–619.

45. Fleiss JL, Davies M. Jackknifing functions of multinomial frequencies, with an application to a measure of concordance. *American Journal of Epidemiology* 1982; **115**:841–845.
46. Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the kappa statistics: confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine* 1992; **11**:1511–1519.
47. Bloch DA, Kraemer HC. 2×2 kappa coefficients: measures of agreement or association. *Biometrics* 1989; **45**:269–287.
48. Landis JR, Koch GG. A one-way components of variance model for categorical data. *Biometrics* 1977; **33**: 671–679.
49. Hanley JA. Standard error of the kappa statistic. *Psychological Bulletin* 1987; **102**:315–321.
50. Fisher RA. *Statistical Methods for Research Workers*, 2nd edn. Oliver & Boyd: London, 1928.
51. Efron B. Bootstrap methods: another look at the jackknife. *Annals of Statistics* 1979; **7**:1–26.
52. Flack VF. Confidence intervals for the interrater agreement measure kappa. *Communications in Statistics—Theory and Methods* 1987; **16**:953–968.
53. Fleiss JL, Cicchetti DV. Inference about weighted kappa in the non-null case. *Applied Psychological Measurement* 1978; **2**:113–117.
54. Blackman NJ-N, Koval JJ. Interval estimation for Cohen's kappa as a measure of agreement. *Statistics in Medicine* 2000; **19**:723–741.
55. Borenstein M. Hypothesis testing and effect size estimation in clinical trials. *Annals of Allergy*, *Asthma*, *and Immunology* 1997; **78**:5–16.
56. Kraemer HC. *Evaluating Medical Tests*: *Objective and Quantitative Guidelines*. Sage Publications: Newbury Park, CA, 1992.
57. Collis GM. Kappa, measures of marginal symmetry and intraclass correlations. *Educational and Psychological Measurement* 1985; **45**:55–62.
58. Kraemer HC, Kazdin AE, Offord DR, Kessler RC, Jensen PS, Kupfer DJ. Measuring the potency of a risk factor for clinical or policy significance. *Psychological Methods* 1999; **4**:257–271.
59. Ross DC. Testing patterned hypotheses in multi-way contingency tables using weighted kappa and weighted chi square. *Educational and Psychological Measurement* 1977; **37**:291–307.
60. Cohen J. The cost of dichotomization. *Applied Psychological Measurement* 1983; **7**:249–253.
61. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Wadsworth & Brooks/Cole Advanced Books & Software: Monterey, CA, 1984.
62. Noda AM, Kraemer HC, Yesavage JA, Periyakoil VS. How many raters are needed for a reliable diagnosis? *International Journal of Methods in Psychiatric Research* 2001; **10**:119–125.
63. Kraemer HC. How many raters? Toward the most reliable diagnostic consensus. *Statistics in Medicine* 1992; **11**:317–331.
64. Kraemer HC. What is the 'right' statistical measure of twin concordance (or diagnostic reliability and validity)? *Archives of General Psychiatry* 1997; **54**:1121–1124.
65. Kraemer HC. Statistical issues in assessing comorbidity. *Statistics in Medicine* 1995; **14**:721–733.
66. Tanner MA, Young MA. Modeling agreement among raters. *Journal of the American Statistical Association* 1985; **80**:175–180.

1.4 Survival Models

TUTORIAL IN BIOSTATISTICS
SURVIVAL ANALYSIS IN OBSERVATIONAL STUDIES

KATE BULL[1] AND DAVID J. SPIEGELHALTER*[2]

[1] *Cardiothoracic Unit, Hospital for Sick Children, Great Ormond Street, London WC1N 3JH, U.K.*
[2] *MRC Biostatistics Unit, Institute of Public Health, Forvie Site, Robinson Way, Cambridge CB2 2SR, U.K.*

SUMMARY

Multi-centre databases are making an increasing contribution to medical understanding. While the statistical handling of randomized experimental studies is well documented in the medical literature, the analysis of observational studies requires the addressing of additional important issues relating to the timing of entry to the study and the effect of potential explanatory variables not introduced until after that time. A series of analyses is illustrated on a small data set. The influence of single and multiple explanatory variables on the outcome after a fixed time interval and on survival time until a specific event are examined. The analysis of the effect on survival of factors that only come into play during follow-up is then considered. The aim of each analysis, the choice of data used, the essentials of the methodology, the interpretation of the results and the limitations and underlying assumptions are discussed. It is emphasized that, in contrast to randomized studies, the basis for selection and timing of interventions in observational studies is not precisely specified so that attribution of a survival effect to an intervention must be tentative. A glossary of terms is provided.
 Stat. Med., Vol. 16, 1041–1074 (1997).

(No. of Figures: 7 No. of Tables: 12 No. of Refs: 31)

1. INTRODUCTION

1.1. Background

As multi-centre databases become established,[1,2] more reports relating clinical outcomes to risk factors and time are emerging. Such studies may have a variety of objectives: description of the experience of a set of patients; identification of risk factors; prediction on individuals for decision-making, and, increasingly, standardization ('risk stratification') for comparisons between centres or even between operators. Investigators may also wish to draw conclusions about the efficacy of alternative interventions or clinical strategies, although the dangers of making judgements about the benefits of treatment from the analysis of databases have been well argued.[3]

1.2. Structure of this paper

The statistical handling of randomized experimental studies has been well discussed in tutorial papers, particularly the classic articles by Peto and colleagues.[4,5] Good observational and randomized studies have much in common, but there are vital differences. Most important is that a randomized study focuses all attention on estimating the effect of an intervention, and balance between treatment groups with respect to known and unknown explanatory variables is assured,

* Correspondence to: D. J. Spiegelhalter

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
 ISBN: 0-470-02365-1

apart from the play of chance, by the act of randomization. Thus, any major observed differences in outcomes may be attributed to a causal effect of the intervention. In contrast, the basis for selection and timing of interventions in observational studies is not precisely specified and attributing effect to cause must be tentative. Thus, in circumstances where randomized studies are not feasible, good observational studies not only require rigorous attention to the quality of the data but also call for more sophisticated statistical analysis.[6,7] This paper's purpose is to identify some problems in designing and analysing observational studies to increase the likelihood that valid conclusions are drawn, and to illustrate some statistical techniques that have been found helpful. The paper is particularly directed at numerate physicians and surgeons with access to a personal computer and at least one of the many statistical packages available, but also may be useful to statisticians responsible for summarizing time-related outcome data.

Two themes are developed in parallel. First, we may be interested in the occurrence of an event within a *fixed interval*, say 'death within a year of surgery'. Second, we may wish to analyse occurrence of events over a period, say 'pattern of mortality up to the age of 20'. We first deal with simple descriptions of data within these two circumstances. We then introduce a single potential risk factor and subsequently consider multiple, possibly interrelated risk factors. Using a small data set, for each analysis we define its aim, the choice of data to be used, the essentials of the methodology, a computational guide with specific attention to interpreting the output of statistical packages, and finally in a *caveat* section we consider the inferences offered in the light of the limitations of the analysis and the plausibility of its underlying assumptions. A non-technical glossary of terms is provided as an Appendix.

The main novelty in this paper concerns techniques for dealing with occurrences which arise while a study is in progress. First, in Section 4 we introduce the concept of *late entry*; for example, the incorporation of data on a patient who did not present to the hospital until late childhood into a study summarizing events for a class of patients from birth. Second, in Section 9 we consider *time-dependent variables*, in which a subject changes status in some way during follow-up, perhaps by having an operation performed. Finally, for the most determined readers, we show in Section 10 how all these concepts may be combined within a single statistical analysis. However, in discussion we emphasize the tentative nature of the conclusions to be drawn from such analyses.

Though many of the calculations for the examples can be carried out on a hand calculator, the full data set and all of these examples were analysed on a personal computer using the readily available statistical packages SPSS (SPSS Inc, 1992) and EGRET (Epidemiological GRaphics Estimation and Testing package; Statistics and Epidemiology Research Corporation 1991), though several other packages are available which will accomplish most of these analyses. We include an annotated example of the necessary SPSS commands in Appendix I.

We should make clear that this is not a comprehensive review of the appropriate statistical methodology and its limitations. For more detailed expositions on survival analysis (in increasing order of mathematical difficulty) see Healy,[8] Altman,[9] Clayton and Hills,[10] Fisher and van Belle,[11] Cox and Oakes[12] and Andersen *et al.*[13] Large prospective epidemiological studies such as the Framingham Heart Study[14] have made extensive use of analyses such as those described in this paper, while the general issues of bias in observational studies have been covered in texts on clinical epidemiology such as Sackett *et al.*[15]

2. DATA

Perhaps the most vital issue in the analysis of any clinical material is the *integrity* of the data, within which we include the quality, completeness and relevance of the information collected. No

amount of analytic sophistication will rescue a project that does not feature these properties,[16] but here we shall, perhaps naively, take them for granted.

2.1. General problems of bias

We have already stressed the problem of drawing any causal interpretation of associations found in observational studies, but there are other general problems of bias that, although they can occur in randomized trials, are particularly prevalent in observational studies. Two issues are introduced below; other potential biases associated with specific types of analysis are described later.

(a) *Bias which prejudices external generalizability.* The aim of a study will usually be to derive from an available subset of patients, statements about their patterns of survival which will be generalizable to a wider body of patients with the condition. There are many instances where there must be concern that the subset of patients in the analysis are not representative of patients as a whole (reports from institutions attracting 'difficult' caes, older cases, 'correctable' cases etc.). If a patient group whose spectrum of disease is not broadly typical is analysed and the conclusions are to be 'generalizable', factors which make them atypical must be accounted for in the analysis.[7] There is then some hope that patient-specific explanatory variables derived from the skew subset can be applied to other patients.

(b) *Ascertainment bias.* This occurs if the availability of information about a patient's status is dependent on that status. For example, patients may be discharged to the care of referring physicians. If a letter is received reporting the death of a patient, how is this information to be used? If notification is more likely to follow a death than a report that the patient is alive, use of this follow-up information will produce an unfavourable bias. To avoid this bias entirely requires complete ascertainment of status at a point in time.

2.2. Illustrative data

We shall illustrate the analyses using a subset of 30 cases extracted from a larger set of 218 patients with complex pulmonary atresia collected as the basis of an observational study on the presentation and natural history of this disease.[17] Complex pulmonary atresia is a congenital malformation with very abnormal sources of blood supply to the lungs. This particular condition is remarkable for the variability in the age at which patients present to medical attention. Patients were selected for the subset because details of their presentation and history were illustrative for our purposes, so no conclusions about the condition of complex pulmonary atresia can be inferred from these exercises.

The original data collection entailed obtaining dates from the patient record including those of birth, presentation, first operation, death and the date when the patient was last seen. Dates were entered onto a spreadsheet and a date subtraction facility allowed generation of ages at presentation, first operation and death or last contact. Features observable at presentation were defined, obtained and coded and are here exemplified by the size of the intrapericardial pulmonary arteries (paanat) and sex. The original data with the ages (in days), and appropriate time intervals already prepared, are shown in Table I. For easy reference, patients in Table I have been arranged according to age at presentation. Additional variables have been derived for use in later analyses.

The data are also summarized in Figure 1, which displays the age-interval during which each patient was followed up and the events occurring during this period. For example, we can

Table I. Sample data set

								Derived data							
Patient	agepres	agelast	ageopl	dead	sex	paanat	adfol	follow-up	opfpres	unopage	unop-fpre	pre-opded	hadop	ded-lyrpp	age-presx
1	1	1274	−1	0	0	0	1	1273	−1	1274	1273	0	0	0	0
2	2	123	40	1	0	1	1	121	38	40	38	0	1	1	0
3	2	119	−1	1	1	0	1	117	−1	119	117	1	0	1	0
4	3	120	−1	0	1	0	0	117	−1	120	117	0	0	2	0
5	6	10	−1	1	0	0	1	4	−1	10	4	1	0	1	0
6	6	5415	194	0	0	1	1	5409	188	194	188	0	1	0	0
7	7	3261	1041	0	1	0	1	3254	1034	1041	1034	0	1	0	0
8	8	1819	−1	0	1	0	1	1811	−1	1819	1811	0	0	0	0
9	11	696	−1	0	0	0	1	685	−1	696	685	0	0	0	0
10	13	6415	29	0	1	1	1	6402	16	29	16	0	1	0	0
11	29	3127	144	1	0	0	1	3098	115	144	115	0	1	0	0
12	30	423	47	1	0	0	1	393	17	47	17	0	1	0	0
13	35	5794	−1	0	1	0	1	5759	−1	5794	5759	0	0	0	0
14	45	292	62	1	1	1	1	247	17	62	17	0	1	1	0
15	54	68	−1	1	1	0	1	14	−1	68	14	1	0	1	0
16	58	1849	−1	1	0	0	1	1791	−1	1849	1791	1	0	0	0
17	68	343	−1	1	1	0	1	275	−1	343	275	1	0	1	0
18	109	3276	1294	0	0	0	1	3167	1185	1294	1185	0	1	0	0
19	119	207	207	1	0	1	1	88	88	207	88	0	1	1	0
20	121	1430	123	0	1	0	1	1309	2	123	2	0	1	0	0
21	231	308	237	1	0	1	1	77	6	237	6	0	1	1	0
22	258	347	−1	0	0	0	0	89	−1	347	89	0	0	2	0
23	349	3355	383	0	0	0	1	3006	34	383	34	0	1	0	0
24	369	3351	2826	1	0	1	1	2982	2457	2826	2457	0	1	0	1
25	437	547	441	0	0	0	0	110	4	441	4	0	1	2	1
26	771	3834	868	0	0	0	1	3063	97	868	97	0	1	0	1
27	1285	7209	−1	0	1	0	1	5924	−1	7209	5924	0	0	0	1
28	2455	3555	3532	1	1	1	1	1100	1077	3532	1077	0	1	0	1
29	5161	5354	5353	1	1	1	1	193	192	5353	192	0	1	1	1
30	5497	5639	5633	1	0	1	1	142	136	5633	136	0	1	1	1

Ages and time intervals expressed in days

agepres	age at presentation
agelast	age last seen alive (if alive) or age at death (if dead)
ageopl	age at first operation (− 1 for no operation to date)
dead	no = 0, yes = 1
sex	male = 0, female = 1
paanat	size of intrapericardial pulmonary arteries at presentation: absent or tiny = 0, normal or near normal = 1
adfol	adequate follow-up. Study closed less than 1 year since presentation = 0, study closed at least 1 year since presentation = 1
followup	duration of follow-up (agelast-agepres)
opfpres	interval from presentation to first operation (− 1 if no operation to date)
unopage	follow-up before first operation (ageopl − agepres) if operated, (lastage − agepres) if unoperated
unopfpre	interval from presentation to first operation (if operated) or to age last seen (if unoperated)
preopded	death before any operation: 0 = no, 1 = yes
hadop	had an operation 0 = no, 1 = yes
dedlyrpp	died within 1 year of presentation. 0 = no, 1 = yes, 2 = not applicable (i.e. adfol = 0)
agepresx	age at presentation grouped: 0 = less than 365 days, 1 = older than 365 days

immediately see that patient 1 presented soon after birth with normal size pulmonary arteries (paanat = 1) and is still alive without operation aged 3, while patient 10 had an operation soon after presentation and is still being followed up aged 16. In contrast, patients 29 and 30 did not present until their teens, and then both soon had an operation which they did not survive. Figure 2 shows the identical data, but with elapsed time being measured from presentation rather than birth.

3. SIMPLE DESCRIPTION OF OUTCOMES AT A FIXED TIME INTERVAL

The most straightforward studies concern 'yes/no' outcomes within a fixed time interval after some event; a common example is reporting of early post-operative mortality.

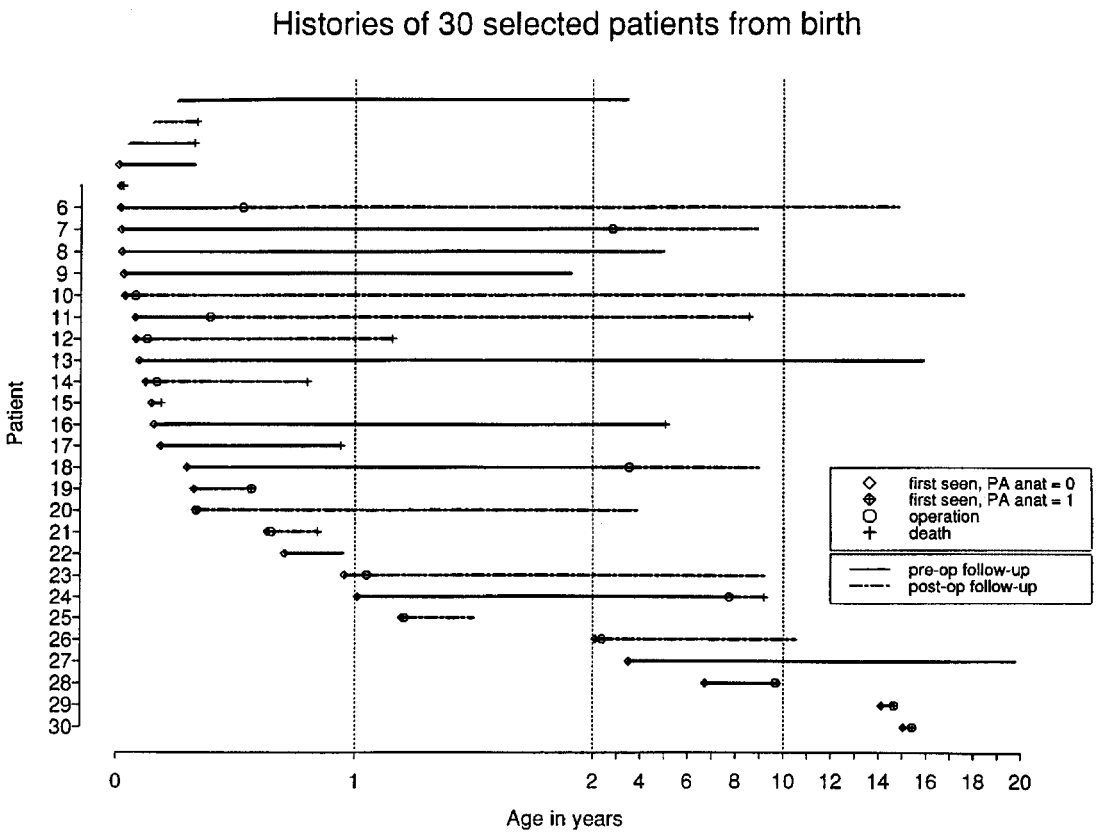


Figure 1. Summary of data showing period of observation of each patient from birth (note the change in scale of the time-axis at 2 years)

3.1. Analysis specification

3.1.1. Inclusion criteria

To be included in such studies, all patients will have had the 'event' referred to (for example, an operation) and in addition, all patients will have been followed (or, if dead, could have been followed) throughout the time interval in question. Examples to be used in our analyses include variables identifying follow-up for at least a year (adfol), and that a patient had an operation at some stage (hadop).

3.1.2. Outcomes (also known as events, responses or dependent variables)

These will include death (perhaps from a particular cause) and possibly intermediate events such as receiving definitive surgery. Examples from our data set include dead and dead within one year of presentation (dedlyrpp).

3.2. Worked example: proportion dying within one year of presentation

We illustrate, using our small data set, the proportion of patients who die within one year of presentation with complex pulmonary atresia (the interval between 0 and 1) in Figure 2. Table II shows the analysis specification using the variable names from Table I.

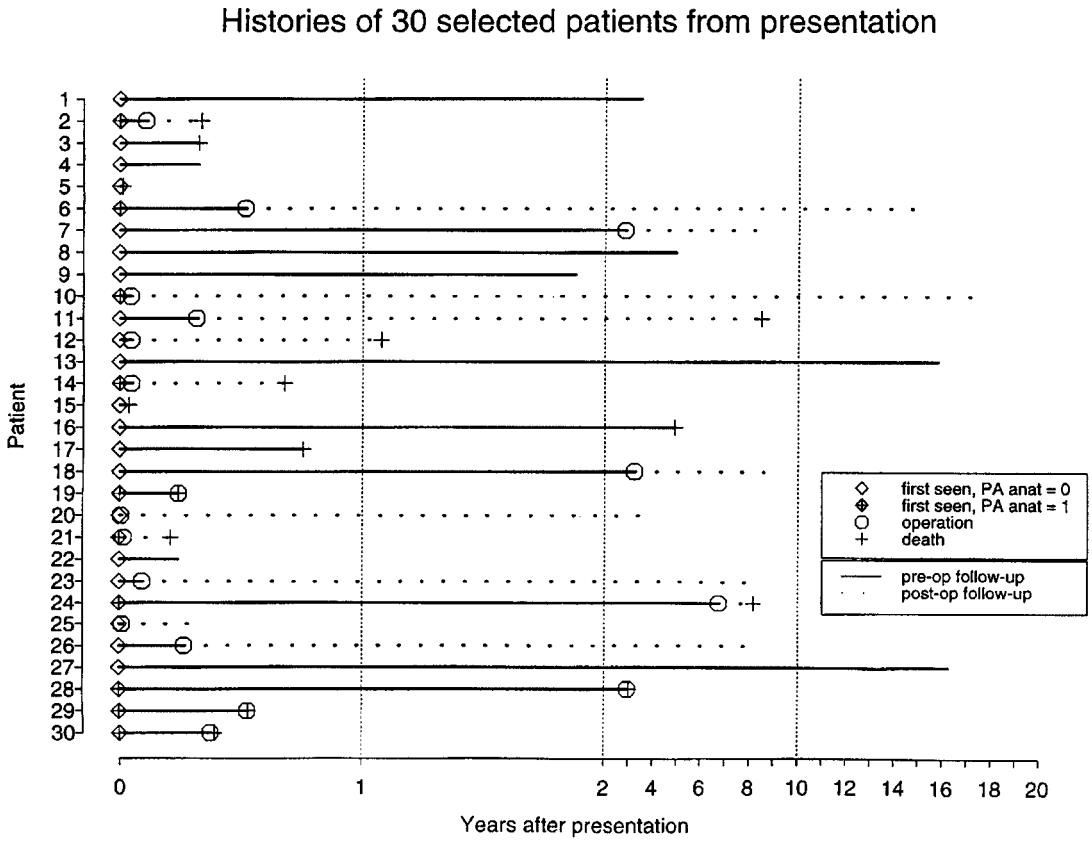


Figure 2. Summary of data showing period of observation of each patient from the time of their presentation (note the change in scale of the time-axis at 2 years)

Table II. Patients dying within one year of presentation

Question:	proportion of patients dying within one year of presentation			
Analysis specification:				
inclusion criteria	patients who have been followed up at least 1 year and patients who (if dead) could have been followed up at least 1 year			adfol = 1 (dedlyrpp ne 2)
outcomes	death within 1 year of presentation			dedlyrpp
Output:				
Dead	Alive	Total n	Proportion dying p	Odds on dying $p/(1-p)$	95% CL on proportion dying $p \pm 1{\cdot}96\sqrt{[p(1-p)/n]}$
10	17	27	10/27 = 0·37	10/17 = 0·59	0·19–0·55

We note that patients 4, 22 and 25 are excluded, being alive but with less than a year of follow-up since presentation (adfol = 0); all the deaths were more than a year before the time of analysis (adfol = 1) and so could have been followed up for a year had they not died.

Table II also presents simple descriptions of the results. An observed proportion p = (number dying/total number) can take on values 0 to 1; it can be transformed into an odds scale $p/(1-p)$ = (the number dying/the number surviving), which can take on values from 0 to infinity. The odds may appear a somewhat unintuitive measure of risk compared to the proportion of events; however, as we shall see later, the odds provide a basis for handling multiple explanatory variables simultaneously and makes a link to the analysis of full survival data. The 95 per cent confidence limits (95 per cent CL) for the true underlying mortality rate were calculated using the standard normal approximation $p \pm 1{\cdot}96\sqrt{\{p(1-p)/n\}}$: more precise estimates[18] are appropriate for smaller number of events (in particular when no events occur) and may be obtained in some statistical packages.

3.3. Caveats and inferences

Generalizations following this simple analysis depend crucially on the *cohort* (the group of patients being followed up) being representative of the overall class of patients of interest. For example, it may be inappropriate to compare such crude mortality rates between hospitals with different referral populations without using the kind of adjustment techniques to be discussed later.

4. BASIC SURVIVAL ANALYSIS (WITHOUT EXPLANATORY VARIABLES)

4.1. Introduction

The previous analysis is limited in two ways; first, it only considers whether an event has occurred by a particular time, and second, it only includes patients who have been, or could have been, followed throughout that time. In contrast, a survival analysis uses information from the whole follow-up period and all patients can contribute information during their time under surveillance.

Generally, the aim of a survival analysis is to use the data available to provide estimates of the probability of surviving to (or being free of the event in question at) different times, this relationship being expressed as the *survival function.* A graph of the survival function provides the most appealing summary of the time-related information.

Suppose we wish to provide a summary of the pattern of survival of patients with complex pulmonary atresia from the time of presentation. If everyone had been meticulously followed from presentation, and if everyone had presented more than 20 years previously, then estimating the survival function up to age 20 would be trivial: we could simply count the proportion who still survived at each age. However, in practice the more recent patients have not been followed-up for so long and so should only contribute to the estimation of survival up to their current age. Standard analyses, demonstrated below, deal with this problem which is known as *censoring* – the loss to follow-up of patients from causes other than the event of interest.

4.2. Analysis specification

In addition to specifying the *inclusion criteria* and *outcomes*, for a survival analysis we also need to define the following terms.

4.2.1. Time origin

This specifies when the 'clock starts' and derives directly from the question posed. In a randomized study the reference for all follow-up is the point of randomization. In observational studies we might wish to ask questions about survival from 'birth' (or even conception) when analysing

natural history, from 'presentation' when studying an acute illness, or 'operation' when investigating the effects of alternative interventions. Figure 1 illustrates the data with time origin at birth and age along the horizontal axis; Figure 2 presents the same data with the time origin at presentation and with years after presentation on the horizontal axis.

4.2.2. Entry to study

Analysis of a randomized study is straightforward because the point of randomization is clearly both the time-origin of the study and the point of entry of every patient to the study. However, in an observational study, the time origin of the study and the beginning of the period of observation of the patient may not coincide (the patient may come under observation before or after the time origin of the study) and so decisions about what represents 'entry to the study' may require more thought. 'Late entry' describes situations where for some or all patients there is a delay between the time origin of the study (specified by the scientific question posed) and the entry time (limited by the data available). See Section 4.10(b).

4.2.3. Withdrawal from the study (censoring)

There are a number of reasons why follow-up of a subject may cease before the event of interest has occurred, although this will usually be simply due to the selected date for the close of the study being reached. The greatest care is required when the current status of the subject is unavailable, since we need to be able to assume that their loss to the study is unrelated to their underlying risk (an assumption known as *non-informative* censoring), since biased results would arise from systematic withdrawal of either high or low risk patients (see Section 4.10(a)).

4.2.4. Survival time

Specification of the time origin of the study, the outcome of interest and the censoring rules determines the *survival time* within the study for each patient. This is the interval between the time origin and either the occurrence of the outcome or censoring. Examples from our data set include the age when last seen (agelast) and the interval between presentation and last contact (followup).

4.2.5. Period of observation

Specificatioin of the entry time and the outcome and censoring rules determine the extremes of the *period of observation* within the study for each patient. This is the interval between the entry time and either the occurrence of the outcome or censoring. In situations with late entry, this may be shorter than the survival time.

4.3. Non-parametric survival functions

We shall first illustrate how *censoring* can be accommodated within the Kaplan–Meier (K–M) procedure,[12] this is known as a *non-parametric* way of estimating a survival function since it makes no assumptions about the shape of the underlying survival curve (it does not assume that it can be summarized mathematically by a limited number of parameters).

4.4. Worked example: survival from presentation (corresponding to Figure 2)

In this analysis the time origin of the study and the point of entry of every patient to the study is the same, and so survival time and period of observation are identical. This will generally be true

Table III. Estimation of survival from presentation

Question:	survival experience of all patients from presentation	
Analysis specification:		
inclusion criteria	all patients	
outcome	death	dead
time origin	presentation	
entry time	presentation	0
censoring rule	withdrawn at end of study	
survival time	time from presentation until death or censored	followup
period of observation	presentation until death or censored	0 to followup

Output:

Patient	Event time	at risk	K–M survival estimate	95% CL on survival estimate	
5	4	30	0·97	0·79	0·99
15	14	29	0·93	0·76	0·98
21	77	28	0·90	0·72	0·97
19	88	27	0·87	0·68	0·95
3	117	24	0·83	0·64	0·93
2	121	22	0·79	0·60	0·90
30	142	21	0·76	0·55	0·88
29	193	20	0·72	0·51	0·85
14	247	19	0·68	0·47	0·82
17	275	18	0·64	0·44	0·79
12	393	17	0·60	0·40	0·76
28	1100	15	0·56	0·36	0·73
16	1791	12	0·52	0·31	0·69
24	2982	10	0·47	0·26	0·64
11	3098	7	0·40	0·20	0·60

for studies of post-operative survival (time origin at operation) or for randomized trials (time-origin at the point of randomization), so a similar analysis will be appropriate in these situations.

Table III shows the analysis specification for this example. The K–M procedure estimates the instantaneous risk of death at any particular time as the ratio of the number who died at that time to the number in the current 'risk set', which is defined to be the set of individuals currently at risk of experiencing the outcome of interest. Hence at the first death (of patient 5) 4 days after presentation, there were 30 in the risk set and hence the risk of death on the 4th day after presentation is estimated to be $1/30 = 0{\cdot}033$. Thus the chance of surviving past 4 days after presentation is estimated to be $1 - 1/30 = 0{\cdot}967$, with 95 per cent confidence limits of 0·79 to 0·99; these limits are best not obtained from an estimated standard error, but from standard formulae[12] that provide assymetric intervals. Fourteen days after presentation, patient 15 dies with a risk set now comprising 29 individuals; the chance of surviving the 14th day after presentation is therefore estimated to be $(1 - 1/29) = 0{\cdot}965$, and thus the estimated cumulative probability of surviving past 14 days becomes $0{\cdot}967 \times 0{\cdot}965 = 0{\cdot}933$ with 95 per cent confidence limits of 0·76 to 0·98, and so on. Results in Table III have been rounded to two figures to reflect the general reporting as 'percentage survival'. Figure 3 displays the estimated survival curve in the conventional 'step' manner.

In mathematical notation, suppose there are r_k subjects in the risk set at the time of the kth distinct time of death t_k, and that at that time there are f_k deaths. Then the estimated survival

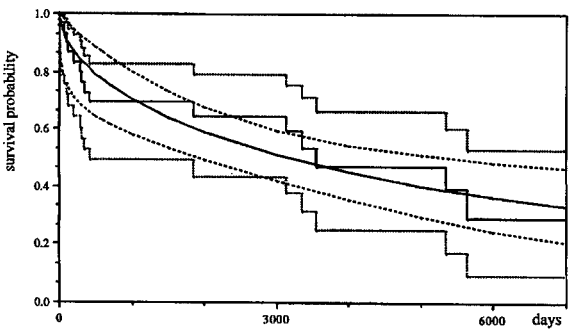


Figure 3. Kaplan–Meier and Weibull estimates of survival from the time of presentation, each with 95 per cent confidence limits

probability until the time t_K is given by

$$p_K = \left(1 - \frac{f_1}{r_1}\right)\left(1 - \frac{f_2}{r_2}\right) \cdots \left(1 - \frac{f_K}{r_K}\right).$$

4.5. Non-parametric survival functions: effect of late entry

If a summary of the survival of patients with complex pulmonary atresia from birth (see Figure 1) is required, a second problem emerges since patients do not come under observation until presentation; this is a general issue in attempting to model the natural history of disease.[19] How we handle *left-truncation* or *late entry* depends on our understanding of the reporting of events for the patients under study. If, for example, we are sure that an event occurring at any point of their life would be reported to us, whether or not the patient was under active follow-up, then we could assume surveillance started at birth and individuals would not enter the cohort late. Such a situation is only plausible in well-defined communities with efficient notification procedures, and as such is rarely an appropriate assumption. Otherwise, avoidance of bias requires that we only include information gathered from patients while they are actively under surveillance.

Because we would usually assume that if an event happened to a patient before they 'presented' we would have been unaware of it, patients should not contribute to our estimate of survival until their age at presentation. An extreme example occurs when we only have information about adult patients – we cannot use them to say anything about survival in childhood. However, just as in right-censoring, we would like to assume *non-informative* late entry,[20] meaning that individuals who present at a certain age are essentially comparable with those of the same age already being followed up; the reasonableness of this assumption is discussed in Section 4.10(a).

4.6. Worked example of survival with late entry: survival from birth (corresponding to Figure 1)

We shall summarize the whole survival experience from birth of all patients with complex pulmonary atresia based on our small selected data set (Figure 1).

Table IV sets out to compare overall survival estimated when all patients are allowed to contribute to the risk set from birth with the curve prepared only allowing patients to contribute to the risk set from presentation; the first is as if a patient's period of observation as illustrated in Figure 1 was extrapolated backwards to birth (appropriate under the optimistic assumption that all events on these patients since birth would have been reported to us). In each case the formula from Section 4.4 is used, with the appropriate size of risk set. Figure 4 plots the estimated survival

Table IV. Survival estimates: entry time 'birth' contrasted to entry time 'presentation'

Question:	whole survival experience of patients in the dataset	
Analysis specification:		
inclusion criteria	all patients	
outcome	death	dead
time origin	birth	
entry time	entry: (a) at birth	0
	(b) at presentation	agepres
censoring rule	withdrawn at end of study	
survival time	birth to death or censored	agelast
period of observation	entry to death or censored	(a) 0 to agelast
		(b) agepres to agelast

Output:

Patient	Event time	(a) from birth at risk	(a) from birth K–M	(b) from presentation at risk	(b) from presentation K–M
5	10	30	0·97	8	0·88
15	68	29	0·93	16	0·82
3	119	28	0·90	17	0·77
2	123	26	0·87	16	0·72
19	207	25	0·83	15	0·68
14	292	24	0·80	16	0·63
21	308	23	0·76	15	0·59
17	343	22	0·73	14	0·55
12	423	20	0·69	14	0·51
16	1849	14	0·64	11	0·46
11	3127	13	0·59	11	0·42
42	3351	10	0·53	8	0·37
28	3555	8	0·47	6	0·31
29	5354	6	0·39	5	0·25
30	5639	4	0·29	4	0·18

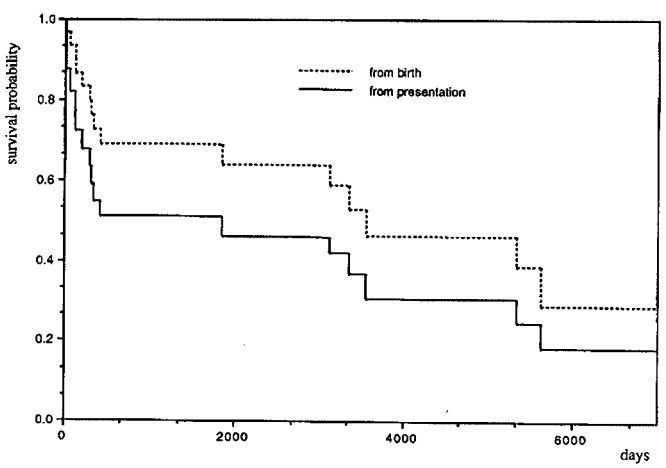


Figure 4. Kaplan–Meier estimates of survival from birth: time origin birth contrasted with time origin at presentation

functions; the survival function assuming entry to the cohort at birth is considerably more optimistic than that generated from the data using only information emerging during the time patients were then being followed up (entry to the cohort at presentation). The pattern of 'drops' reflects the fact that the only difference between the two functions is due to the size of the risk set at each time of death. For example, the first death occurred at age 10 days. We know that 30 children were alive at that age, but only 8 of them were actually being followed-up within this study. The choice of the risk set (with 8 or 30 individuals) determines the size of the denominator in the calculation, and a larger denominator will always decrease the apparent risk.

4.7. Parametric survival functions

Non-parametric techniques use the data to 'draw' the survival function directly; the methodology will conform to any pattern of survival but the survival function proceeds by downward steps which do not reflect the usual perception of an underlying continuity in nature. *Parametric* survival functions reflect both the data and some assumptions about it, including an underlying continuity. The assumptions are embodied in parameters which are themselves estimated from the data; the resulting survival function is thus a mathematical equation which describes a smooth curve. Simple parametric functions include the exponential (in which a single parameter characterizes the death rate which is assumed constant), Weibull (with two parameters allowing the death rate to either increase or decrease with time) or a variety of other forms.[12]

Figure 3 shows a fitted Weibull curve for survival from the time of presentation and its 95 per cent confidence limits (see below for details of this fitted curve). The comparable K–M function is shown with its confidence limits. We note that the confidence limits for the parametric curve are tighter than for the non-parametric curve; this additional precision has been obtained at a cost of greater assumptions which may lead to additional bias. For example, the Weibull curve does not appear to have sufficiently captured the high early mortality followed by the rapid reduction in risk for patients surviving a year from presentation. More complex parametric models are available which adapt to the different survival patterns for early and late mortality;[21] such models have more free parameters which allow greater adaptation to observed patterns and tend to produce more precise estimates than a non-parametric analysis. Such complex parametric models do have the disadvantage that the survival curve at a particular time may be substantially influenced by temporally distant events, and in particular that it may be tempting to extrapolate the curve beyond the region in which the evidence is strong.

4.8. Hazard functions

While survival curves express the *cumulative* effect of the risks faced by an individual, it is often both more convenient and interpretable to work directly in terms of *hazard* at a point in time (the risk of an event occurring per unit of time elapsed, given that the individual has survived up to that time). The *hazard function* expresses the hazard as it changes over time and contains exactly the same information as the survival function but in terms of its rate of change; where the survival curve is falling fast, hazard is high, while if the survival curve is flat the hazard is zero. This equivalence allows the hazard function to be derived by a mathematical transformation of the survival function (see following), as in Figure 5 where a smooth parametric Weibull survival function curve transforms to a smooth hazard function. An identical transformation can be applied to the survival function generated by K–M methodology, though the hazard function then appears as a series of spikes because the K–M survival function falls by discrete steps, and generally has to be smoothed to produce a reasonable plot. In the next section we show how to obtain a slightly different direct estimate of the average hazard within a time window.

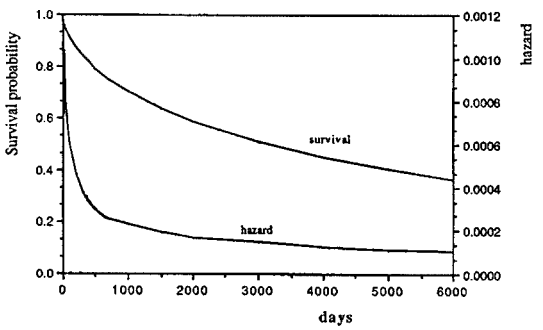


Figure 5. Weibull survival function and hazard function over the first year from presentation

4.9. Computation

Kaplan–Meier estimates are available in most packages, but rarely is adjustment for late entry allowed; exceptions include EGRET. Alternatively, late entry calculations as shown in Table IV could be carried out by hand or on a spreadsheet. Exponential and Weibull parametric models can be fitted in EGRET, but, as is common with most programs for parametric survival analysis, late entry cannot be handled.

If we denote the survival and hazard functions at time t by $S(t)$ and $h(t)$, respectively, then it can be shown that the hazard equals minus the derivative of the natural logarithm of the survival function, or $h(t) = -\,\mathrm{d}\log S(t)/\mathrm{d}t$. For an exponential distribution, $S(t) = \exp(-bt)$, (where the mean survival is $1/b$) and hence $h(t) = b$ reflecting the equivalent assumption of a constant hazard. For a Weibull distribution we assume $S(t) = \exp(-(bt)^k)$, and hence $h(t) = kb(bt)^{k-1}$, which means that the hazard function may be increasing ($k > 1$) or decreasing with time ($k < 1$). The fitted curve in Figure 3 has parameters $b = 0{\cdot}00021$ and $k = 0{\cdot}46$.

4.10. Inferences and Caveats

(a) *Informative right-censoring.* We have previously mentioned the necessity of assuming that being withdrawn from follow-up is unrelated to the current hazard of death ('non-informative censoring'). This will generally not be a problem if censoring occurs due to the end of the study period, or because of loss to follow-up due to an event unrelated to the underlying risk; events such as emigration or accidental death would usually be considered as non-informative, although this may not be strictly appropriate. In other circumstances censoring may well be informative. Suppose a researcher wished to estimate the 'natural history' of a disease, and censored patients at a definitive operation. This would only be non-informative if patients were selected for operation on the basis of factors unrelated to their current risk of death, which would rarely be the case.

(b) *Informative late entry.* As with censoring, we wish to assume that late arrivals into the study are at equal risk to those already under surveillance ('non-informative late entry'). This will not be a problem if those starting to be followed-up are similar to those already under surveillance, which would occur if, say, new patients were identified through a broadening of the scope of the study. Often, however, new patients will have been referred because of a worsening condition, or, conversely, because they seem a good candidate for a definitive intervention. In either case, age at presentation to a secondary referral institution is likely to be an important predictive variable, possibly being a proxy for the current severity of illness. There is a simple test to examine the independence of time of

entry and survival time,[22] although an alternative means of testing this assumption is by including age at entry into a Cox regression analysis (Section 9) and checking it has no effect upon subsequent risk.

In addition to the standard requirements for generalizability, we now need to address assumptions concerning censoring and late entry. It is apparent that each increment in complexity of analysis, while bringing with it a more realistic representation of the realities underlying the data, require associated judgements on conformity to broad assumptions. Even if we are happy about making such assumptions, it is clear that the simple descriptive analyses shown do not formally explore factors that may influence the outcome; we now need to introduce methods for making comparisons between groups of patients.

5. OUTCOMES AT A FIXED TIME INTERVAL: ONE FACTOR AT A TIME

The simple description of events within a fixed time interval can be readily extended and becomes clinically more useful when the influence of possible explanatory variables can be explored.

5.1. Analysis specification

In randomized trials all explanatory variables (also known as *risk factors*, *covariates*, *predictors*, or *independent variables*) are defined at the point of randomization and are guaranteed (apart from chance variation) to be balanced between treatment groups by the act of randomization. Typically these variables will include age, morphology, clinical status and centre. In the absence of randomization, treatment groups will not be balanced with respect to such variables and hence it is vital that all known explanatory variables are recorded so that the analysis can attempt to adjust for them. Examples from our data set include age at presentation (agepres), gender (sex) and pulmonary artery anatomy (paanat).

5.2. Worked example: deaths within one year of presentation

Here we explore the influence of two variables pulmonary artery anatomy (paanat) and age at presentation, grouped into less than or greater than one year (agepresx) as predictors of death within one year of presentation; both explanatory variables were observable at presentation. In general, variables may be discrete or continuous, although in an exploratory analysis it is generally helpful to group any continuous quantity into discrete categories (such as agepresx); these are generally called *factors*. For each category of factor explored, Table V begins by showing the proportions and percentages dying within a year of presentation.

For each factor explored in this way, a baseline category is identified relative to which comparisons are made. This baseline category is generally the lowest risk or the most common category; here paanat = 0 and agepresx = 0 are used. The odds ratio, relative to baseline, is then calculated for each non-baseline category: these odds ratios are known as *unadjusted* or *simple odds ratios* since they only take into account the association between single factors and outcome. For example, the odds ratio for paanat = 1, relative to paanat = 0, is (6/4)/(4/13) = 4·88.

5.3. Computation

All statistical packages should be able to cross-tabulate a categorical factor against an outcome measure and calculate p-values using simple chi-squared tests. Confidence intervals and significance levels for the odds ratio are generally obtainable from statistical packages. In the example there is an excess odds on death, associated with having normal or near-normal pulmonary artery

Table V. Univariate and multivariate analysis of outcomes after a fixed time interval

Question:		predictors of death within one year of presentation								
Analysis specification:										
inclusion criteria		patients who have been followed up at least 1 year *and* patients who (if dead) could have been followed up at least 1 year					adfol = 1 (dedlyrpp ne 2)			
outcomes		death within 1 year of presentation					dedlyrpp			
explanatory variables		pa anatomy					paanat			
		age at presentation (grouped)					agepresx			
Output:										
Factor	Category	Mortality	%	Odds on death	Univariate analysis			Multivariate analysis		
					odds relative to baseline	95% CL on odds relative to baseline	*p*-value	odds relative to baseline	95% CL on odds relative to baseline	*p*-value
paanat	0	4/17	24%	4/13	1·00			1·00		
	1	6/10	60%	6/4	4·88	0·90–26·42	0·07	6·94	1·01–48·03	0·05
agepresx	0(< 1 year)	8/21	38%	8/13	1·00			1·00		
	1(⩾ 1 year)	2/6	33%	2/4	0·81	0·12–5·50	0·83	0·35	0·04–3·43	0·36
						baseline odds		0·34		

size rather than absent or hypoplastic pulmonary arteries, of 4·88 with 95 per cent confidence interval 0·90 to 26·42; this wide interval just includes 1 and hence we cannot strictly exclude the possibility that pulmonary artery size is not associated with a change in mortality within one year of presentation. This is reflected in the *p*-value of 0·07, which states that there is a 7 per cent chance of observing such an extreme odds ratio even if there were no change in risk. (In general we note that a 95 per cent confidence interval just excluding 1 is essentially the same as a chi-squared test of association rejecting at the 5 per cent level the null hypothesis of no difference from baseline risk.)

5.4. Caveats and inferences

Two concerns with explanatory variables can be identified. First, for results to be generalizable we must be sure that the variables are measured similarly in other contexts; this is easy for precise factors such as agepresx but may be more contentious when subjective judgements about morphology are involved (paanat). Second, looking at one factor at a time can be misleading and we need to consider techniques for examining multiple factors simultaneously (see Section 7.1).

6. SURVIVAL WITH ONE FIXED EXPLANATORY VARIABLE

It is clear that K–M estimated survival functions may easily be calculated for two categories of patient.

6.1. Worked example: survival in different risk groups

Here we consider the example of patients with paanat 0 and 1, using 'presentation' as the time origin. The analysis specification is shown in Table VI, and in the output, the event times correspond with the risk set in Table III broken down into pa anatomy categories. The survival functions are plotted in Figure 6.

Table VI. Example: survival estimates and direct estimates of hazard in the first year and 1–10 years after presentation according to pulmonary artery anatomy

Question:	survival from presentation according to pa anatomy	
Analysis specification:		
inclusion criteria	all patients	
outcome	death	dead
time origin	presentation	
entry time	presentation	
censoring rule	withdrawn at end of study	
survival time	time from presentation until death or censored	followup
period of observation	presentation until death or censored	0 to followup
explanatory variables	pa anatomy	paanat

Output:

Patient	Event time	paanat = 0					paanat = 1				
		at risk	events	K–M	instantaneous hazard/day	estimated hazard/year	at risk	events	K–M	instantaneous hazard/day	estimated hazard/year
to 1 year after presentation											
5	4	20	1	0·950	1/20 = 0·050		10	0	1·00		
15	14	19	1	0·900	1/19 = 0·053		10	0	1·00		
21	77	19	0	0·900			10	1	0·900	1/10 = 0·1	
19	88	19	0	0·900			9	1	0·800	1/9 = 0·111	(0·1 + 0·111 + 0·125
3	117	16	1	0·844	1/16 = 0·063	(0·050 + 0·053	9	0	0·800		+ 0·143 + 0·167
2	121	16	0	0·844		+ 0·063 + 0·071)	8	1	0·700	1/8 = 0·125	+ 0·2)
30	142	16	0	0·844		**= 0·237**	7	1	0·600	1/7 = 0·143	**= 0·846**
29	193	16	0	0·844			6	1	0·500	1/6 = 0·167	
14	247	16	0	0·844		(SE 0·123)	5	1	0·400	1/5 = 0·2	(SE 0·387)
17	275	14	1	0·783	1/14 = 0·071		5	0	0·400		
1 to 10 years after presentation											
12	393	13	1	0·726	1/13 = 0·077		5	0	0·400		
28	1100	13	0	0·726		(0·077 + 0·111	4	1	0·300	1/4 = 0·25	(0·25 + 0·333)/9
16	1791	9	1	0·643	1/9 = 0·111	+ 0·2)/9	4	0	0·300		**= 0·065**
24	2982	9	0	0·643		**= 0·043**	3	1	0·200	1/3 = 0·333	
11	3098	5	1	0·514	1/5 = 0·2	(SE 0·029)	3	0	0·200		(SE 0·055)

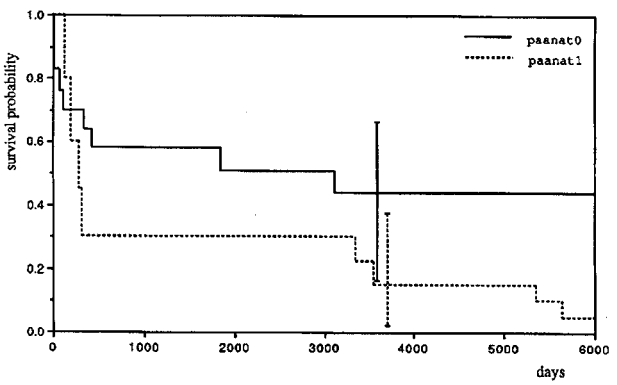


Figure 6. Kaplan–Meier survival estimates according to pulmonary artery anatomy, each with a 95 per cent confidence limit at 3650 days

6.2. Calculation of hazard and its standard error

As described in Section 4.8, general comparison of survival experience may be approached in terms of hazard. Estimates of the instantaneous hazard per day can be derived directly from the observed event rate as in Table VI and from these the average or 'smoothed' hazard over an

interval may be derived. For example, for paanat = 0, we may estimate the instantaneous risk or hazard on the 4th day after presentation, when there was one death from a risk set size 20, to be 1/20 = 0·05. Each subsequent event contributes to an estimate of the hazard at that time, and accumulating these provides an estimated cumulative hazard over a specific period. In this example we have estimated the annual hazard.

The appropriate formulae are provided, for example, by Cox and Oakes,[12] p. 56. Remembering the notation introduced in Section 4.4 (r_k is the size of the risk set in which f_k deaths occur), and consider a time period of length T in which the first K distinct times of death occur. Then the cumulative hazard over this period may be estimated by

$$\frac{f_1}{r_1} + \frac{f_2}{r_2} + \cdots + \frac{f_K}{r_K}$$

with estimated standard error

$$\sqrt{\left\{\frac{f_1}{r_1(r_1 - f_1)} + \frac{f_2}{r_2(r_2 - f_2)} + \cdots + \frac{f_K}{r_K(r_K - f_K)}\right\}}.$$

These formulae can be trivially generalized to any follow-up period, not necessarily starting at time zero.

The average hazard and its standard error may be obtained by dividing each of these quantities by the length of the period T, as shown in Table VI. However, this is a purely descriptive quantity; if it is believed that the hazard was constant over the entire period then an exponential survival distribution should be fitted, with hazard rate estimated by the total number of failures divided by the total follow-up time during the period.

Often the actual hazard in a group is not of primary interest, but attention focuses on the ratio of the hazards between the categories of a factor. In our example the estimated hazard ratio is 0·850/0·237 = 3·59 during the first year after presentation and 0·064/0·043 = 1·49 from years 1 to 10 post-presentation. An important question is whether it might be reasonable to assume that the hazard ratio does not depend on the patient's time from presentation, since this would mean we could unambiguously talk of the hazard ratio associated with a particular pulmonary artery anatomy. The assumption that the hazard ratio does not depend on the elapsed time is known as *proportional hazards*, and this rather stringent assumption is fundamental to much of survival analysis.

Though hazard ratios may be estimated directly as in Table VI, in general it is easier to use the Cox regression model described in Section 8, where we also discuss formal tests for proportionality of hazards.

6.3. Comparison of survival between two groups

Comparison between the survival at any chosen time, say 1-year survival, is possible by computing approximate p-values based on the observed survival difference and its estimated standard error. For many other purposes some comparison of the whole survival experience of the two groups will be desirable. This requires the logrank or Cox–Mantel test.

6.4. Worked example: Logrank test for comparing survival between patients with different pa anatomy

The assessment of the statistical significance of the observed difference between K–M survival functions has been discussed in detail by Peto *et al.*,[5] but we show the layout of the data and how these calculations may be carried out in Table VII.

Table VII. Calculation of logrank statistic for comparing survival of groups

Question:		comparison of whole survival experience according to pa anatomy					
Analysis specification:		as Table VI					
Patient	Event time	paanat = 0			paanat = 1		
		at risk	observed	expected	at risk	observed	expected
5	4	20	1	0·667	10	0	0·333
15	14	19	1	0·655	10	0	0·345
21	77	19	0	0·655	10	1	0·345
19	88	19	0	0·678	9	1	0·321
3	117	16	1	0·640	9	0	0·360
2	121	16	0	0·667	8	1	0·333
30	142	16	0	0·696	7	1	0·304
29	193	16	0	0·727	6	1	0·273
14	247	16	0	0·762	5	1	0·238
17	275	14	1	0·737	5	0	0·263
12	393	13	1	0·722	5	0	0·278
28	1100	13	0	0·765	4	1	0·235
16	1791	9	1	0·692	4	0	0·308
24	2982	9	0	0·750	3	1	0·250
11	3098	5	1	0·625	3	0	0·375
			$O_1 = 7$	$E_1 = 10{\cdot}438$		$O_2 = 8$	$E_2 = 4{\cdot}561$

Hazard ratio $(8/4{\cdot}561)/(7/10{\cdot}438) = 2{\cdot}615$

Again we consider survival from presentation for the two groups defined by pulmonary artery anatomy (categories 0 and 1 of paanat). Table VII is similar to Table VI, the *observed* columns indicate the number of deaths in each group at the event time (since there are no 'ties' these are always 1 or 0) and the *expected* columns give the calculated number of deaths that would occur in each group were there no excess risk for either group; for example, at the time of the first death there were 20 at risk with small pulmonary arteries (paanat = 0) and 10 with normal size pulmonary arteries (paanat = 1), so if there were no difference between the two groups we would expect 0·67 of a death in the first and 0·33 of a death in the second group. (It may seem somewhat strange to obtain such fractions of deaths but it is their total that is important.) We then sum the observed and expected columns to give the totals denoted O_1, E_1, O_2, E_2 as shown.

If the two groups had identical risk the expected number of deaths would be close to that observed in each group. In fact there appears to be an excess of deaths in the paanat = 1 group. The statistical significance of this excess can be assessed by calculating a test statistic $\chi^2 = (O_1 - E_1)^2/E_1 + (O_2 - E_2)^2/E_2$ which will be approximately distributed as a chi-square statistic with 1 degree of freedom under the null hypothesis that the survival functions in the two groups are identical. (This approximation is conservative in that the calculated p-value may be larger than appropriate.[8]) Our statistic is $\chi^2 = 3{\cdot}72$, and consulting standard tables reveals that there is about a 6 per cent chance of observing such an extreme result if the groups had the same survival, so $p = 0{\cdot}06$. Thus there is some, but not overwhelming, evidence of a difference between the groups. An estimate of an overall hazard ratio is given by $(O_2/E_2)/(O_1/E_1) = 2{\cdot}62$.

It is worth noting that the logrank test can be made to accommodate data sets with late entry; for example, just as a breakdown of Table IV(a) according to pa anatomy has provided Tables VI

and VIII, Table IV(b) could provide a comparison of overall survival since birth with entry to the study at presentation.

7. OUTCOMES AT A FIXED TIME: MORE THAN ONE EXPLANATORY FACTOR

7.1. Adjusted odds ratios using logistic regression

Univariate, or *unadjusted*, odds ratios may be misleading if explanatory variables are strongly related to each other; for example, when considering a surgical procedure an apparent association between age and mortality might be explained by the fact that older patients have a more severe form of disease. A possible solution would be to examine whether there is still a relationship with age within each severity category, but with more than a few factors such repeated subdivisions of the data lead to numbers that are too small for meaningful analysis. When explanatory variables are themselves associated, what we are really after is a measure of the association between a factor and the outcome assuming all other measured factors are kept fixed. Logistic regression allows the required *adjusted* odds ratios for multiple factors to be estimated simultaneously, assuming such odds ratios are independent of underlying risk and the values other factors take on.

Table V shows that for our simple example, the adjusted odds ratios are slightly different from the unadjusted. The odds on mortality for a patient with paanat = 1 relative to a patient with small pulmonary arteries (paanat = 0), allowing for age at presentation (agepresx) staying constant, is now 6·94, and the 95 per cent confidence limits for paanat now exclude 1. Table V also shows a 'baseline odds' on mortality for an imaginary patient whose factors are all fixed at their baseline categories; thus a patient with paanat = 0 and agepresx = 0 has estimated odds of 0·34 on death within a year of presentation, which translates to an estimated probability of 0·34/1·34 = 0·25 or 25 per cent. (Since odds = probability/(1 − probability), we can invert the relationship to give probability = odds/(1 + odds)). By multiplying this baseline odds by the adjusted odds ratios for observed categories of explanatory variables for a specific patient, we may obtain their estimated odds on death.

In notation, we can let d_0 be the baseline odds, and d_i be the odds ratio associated with the observed category of the ith factor. If I factors are taken into account, the final odds is given by

$$\frac{p}{1-p} = d_0 \times d_1 \times \cdots \times d_I.$$

For example, a patient with both normal size pulmonary arteries (paanat = 1) and older age at presentation (agepresx = 1) would have an estimated odds on dying within one year of presentation of 0·34 × 6·94 × 0·35 = 0·83, or equivalently an estimated probability of dying in that time frame of 0·83/1·83 = 0·45 or 45 per cent. The simplicity of this calculation demonstrates why working in odds ratios is advantageous when dealing with multiple explanatory variables.

7.2. Computation

Most packages will handle unadjusted and adjusted odds ratio estimation within a logistic regression framework. Care is required in handling factors with more than two categories; a variable taking on, say, values 0, 1, 2, 3 will be handled as a continuous variable by default with the implication that the odds ratio between category 0 and 1 is the same as that between categories 1 and 2 and so on. If such a specific relationship is not intended, the categorical nature of the variable must be acknowledged for appropriate analysis. If the software allows, the categorical nature can simply be 'declared', otherwise a series of (0, 1) variables, one for each

Table VIII. Logistic regression output in terms of regression coefficients (adjusted only)

Question:		as for Table V							
Analysis specification:		as for Table V							
Alternative output:									
Factor	Category	B	SE(B)	p-value	95% CL on B		exp(B) (= odds)	95% CL on odds	
					lower	upper		lower	upper
paanat	0						1·00		
	1	1·94	0·99	0·049	1·94 − (1·96 × 0·99) = 0·0004	1·94 + (1·96 × 0·99) = 3·88	6·95	1·01	48·4
agepresx	0 (< 1 year)						1·00		
	1 (≥ 1 year)	− 1·06	1·17	0·36	− 1·06 − (1·96 × 1·17)	− 1·06 + (1·96 × 1·17)	0·35	0·04	3·42
Constant		− 1·08	0·58						

non-baseline category, must be created to allow the effect of each to be compared to baseline. Packages can differ in the way in which comparisons are made between categories of factors (for example, in SPSS for Windows the above standard coding is known as 'indicator'). We note that for dichotomous variables it is convenient to code the categories as 0 and 1, since it is then irrelevant whether the variable is treated as categorical or continuous.

Some packages express the results of a logistic regression in terms of odds ratios and confidence intervals, similar to Table V. Others may only give the results in terms of estimates and standard errors of individual regression coefficients related to the logarithm of the odds on death; these regression coefficients are simply the natural logarithms of the odds ratios. This relationship is demonstrated by taking natural logarithms of the formula in the previous subsection to give

$$\log\left(\frac{p}{1-p}\right) = B_0 + B_1 + \cdots + B_I$$

where $B_0 = \log d_0$ denotes the baseline log-odds, and B_1 to B_I denote the log odds-ratio $\log d_1$ to $\log d_I$. Table VIII denotes the estimate and standard error of any particular coefficient as B and SE(B), giving an approximate 95 per cent interval for the true coefficient of (B − 1·96 SE(B), B + 1·96 SE(B)), (since +/−1·96 standard errors is a 95 per cent confidence interval assuming the estimator is normally distributed). Then the estimated odds ratio and its confidence interval are obtained by taking exponents (anti-log) of the results for B, giving an estimated odds ratio of exp(B) and 95 per cent confidence intervals of exp(B − 1·96 SE(B)) and exp(B + 1·96 SE(B)). The 95 per cent limits for the baseline odds $d_0 = \exp B_0$ may be obtained from the baseline constant in the same way, and apart from rounding errors the results of Table VIII match those of Table V.

We note that we can calculate the estimated probability of any individual surviving one year post-presentation as in Section 7.1, but using the additive coefficients rather than the multiplicative odds ratios. Thus for a patient with normal pulmonary artery size (paanat = 1) and older age at presentation (agepres = 1) we obtain a total B score (the logarithm of the odds on mortality) of (− 1·08 + 1·94 − 1·06) = − 0·2, and hence the odds are $e^{-0·2} = 0·82$ compared to the 0·83 found before. This equivalent result (apart from rounding errors) shows that logistic regression naturally produces a scoring system that can be used for simple risk stratification of patients. In particular, the estimated mortality probabilities for individual patients may be summed to produce an expected mortality within, say, a centre, and then may be contrasted with the observed number of deaths. The resulting comparison will serve as a fairer basis for audit of centres than naive ranking of raw mortality rates, since some adjustment has been carried out for case mix.[23]

Continuous variables are often grouped into categories and hence turned into factors. However, if kept as a continuous quantity and entered into a logistic regression, odds ratios are

interpreted as the change in odds per unit increase in the variable. It is generally useful to subtract a selected 'baseline' value, often the average in the patient sample, in order to retain the interpretation of d_0 as the odds for a baseline patient.

7.3. Caveats

We have steadily elaborated our analyses throughout this paper in an attempt to provide answers to the scientific questions being posed. Such questions may relate to estimating risks, examining associations, predictions on individuals, comparing centres, and even tentatively exploring the causal effects of interventions. The additional power to answer such questions has come through constructing a *model* for the limited data available, which attempts to provide a representation of the underlying mechanisms through making a series of assumptions. We always need to emphasize that a model is never actually *true*, but may be *useful*. The process of model construction, elaboration and criticism is possibly the most vital part of statistical analysis, although the difficulty of formulating strict rules means that it is often left out of standard statistical texts. There is inevitably a strong element of judgement required, and this is best carried out in close collaboration between statisticians and clinicians.

The data may impose limitations on the number of explanatory variables which can be usefully explored simultaneously. Even with many patients available in the database, the main constraint will relate to the number of *events* on which the logistic regression model bases its estimates. A conservative guideline proposed by Harrell *et al.*[24] is to suggest that if there are fewer than 10 times as many *events* to be predicted as there are *explanatory variables* in the model, the p-values associated with the odds related to each variable may be misleading.

In logistic regression it is assumed that odds ratios for the categories of a factor do not depend on the actual categories observed for other factors, but it is possible to specifically include such *interactions* which would allow, for example, the effect of severity of illness to differ according to the age of the patient. However, since there may be many such possible interactions, their selection should largely be based on clinical judgement.[24]

Many packages provide procedures for automatic selection of variables to be included in a model based on stepwise significance testing. Great care is required with the interpretation of the output from these techniques;[24] many significance tests have been done so neither the p-values nor the fact that a variable has been selected or removed should be taken too literally. It is better that variable selection proceeds on the basis of clinical as well as statistical considerations; in particular, the fact that a variable has an odds ratio that is not significantly different from 1 is not in itself a reason to remove it from the model (this would make the error of assuming that the odds ratio really was 1).

Measurement error in explanatory variables is an important consideration; within-individual variability in the measurement will lead to an underestimate of the true odds ratio. This is sometimes known as 'regression dilution bias'. For example, the use of a single diastolic blood pressure measurement leads to a 60 per cent underestimate of the association of diastolic blood pressure with coronary heart disease,[25] compared with the association that exists with an individual's long-term average diastolic blood pressure.

8. SURVIVAL – MANY FIXED FACTORS

8.1. Cox regression using the whole survival experience

Suppose we wish to simultaneously investigate the influence of pulmonary artery anatomy and the gender of the patient. Two survival curves for patients with paanat 0 and 1 have been shown

Table IX. Example of Cox regression: factors with potential to influence survival

Question:	factors with potential to influence survival?	
Analysis specification:		
inclusion criteria	all patients	
outcome	death	dead
time origin	presentation	
censoring rule	withdrawn at end of study	
survival time	time from presentation until death or censored	followup
entry time	presentation	0
period of observation	presentation to end of follow-up	0 to followup
explanatory variables	pa anatomy	paanat
	gender	sex

Output:

Factor	Category	Univariate analysis				Multivariate analysis			
		Hazard ratio relative to baseline	95% CL on hazard ratio relative to baseline		*p*-value	Hazard ratio relative to baseline	95% CL on hazard ratio relative to baseline		*p*-value
paanat	0	1·00				1			
	1	2·68	0·96	7·42	$p = 0{\cdot}06$	2·69	0·97	7·48	$p = 0{\cdot}06$
sex	0	1·00				1			
	1	0·79	0·28	2·21	$p = 0{\cdot}66$	0·77	0·27	2·17	$p = 0{\cdot}62$

in Figure 6, and, in principle, four curves describing the survival experience of patients with each category of pulmonary artery anatomy in each category of gender could be produced. However, when there are many variables to be explored, the strategy of constantly subdividing the data set to provide comparisons will quickly limit the data available in some subgroups and summarizing the contrasts between many survival curves becomes difficult. In the same way that logistic regression provides a simplifying model that allowed estimation of odds ratios when many factors are being explored at the same time, Cox regression is the technique that provides simultaneous estimates of hazard ratios in the presence of multiple explanatory variables.[1,2] In logistic regression the odds ratio is assumed independent of the underlying baseline odds, and similarly in Cox regression the hazard ratio is assumed independent of the baseline hazard function, which can be of any form. We may express this by the formula

$$\text{hazard ratio at time } t = h_0(t) \times h_1 \times \cdots \times h_i$$

where $h_0(t)$ is the baseline hazard function at time t, and h_i is the hazard ratio associated with the observed category of the ith factor. If a single factor is entered into a Cox regression then unadjusted hazard ratios may be estimated and p-values calculated; these p-values will be essentially equivalent to those obtained using the logrank procedure shown previously (see Section 8.2).

By way of example, we extend the previous survival analysis in Table VI in order to explore two variables (paanat and sex) with potential to influence the survival function from presentation. Table IX provides the results; we draw attention to the strong similarities to the layout of Table V.

The adjusted hazard ratios may be interpreted as follows. Relative to a baseline patient who has small pulmonary arteries (paanat = 0) and male gender (sex = 0), there is no strong evidence that a similar female patient (sex = 1) has more or less risk, although the width of the confidence interval shows considerable uncertainty as to the true effect. There is some evidence of an increase in risk with larger pulmonary arteries (paanat = 1), with the best estimate being nearly a 2·7-fold excess death rate, but again with great uncertainly around this estimate. This analysis assumes that this excess risk persists throughout follow-up.

Just as with logistic regression, this allows an estimate of the increased hazard associated with any configuration of observed explanatory variables. For example, a patient with both paanat = 1 and sex = 1 would have an estimated hazard ratio of $2{\cdot}69 \times 0{\cdot}77 = 2{\cdot}07$ over a baseline patient with both factors in category 0.

8.2. Computation

Cox proportional hazards survival analysis is now available in many packages; as is the case in logistic regression, categorical variables need appropriate handling and baseline categories are chosen explicitly or by default. The output is also very similar to that of logistic regression; in particular, results are often provided in terms of the actual regression coefficients representing $\log h_i$, that have to be exponentially transformed, just as in Section 7.2, to yield estimates and intervals for the hazard ratios h_i. Usually it is possible to produce estimated survival curves for any selected configuration of explanatory variables, and estimates of the underlying hazard function are generally available.

Somewhat confusingly, p-values for individual factors can be obtained by three different methods – 'the likelihood ratio' procedure, the 'score test' and the Wald procedure in which the estimated coefficient divided by its standard error is compared with standard normal tables. Fortunately all three approaches generally give similar answers; our quoted p-values are based on the third approach.

8.3. Caveats

Cox regression is known as a *semi-parametric* procedure in which a parametric model for the relative hazard is overlaid on a non-parametric estimate of underlying hazard. With more data it is possible to carry out formal and informal checks of proportional hazards;[26] here we only consider some basic suggestions. One possibility is to divide the follow-up period into a number of *epochs*, corresponding to, say, early, middle and late mortality, and perform a Cox regression analysis separately within epochs. Comparison should then reveal whether estimated hazard ratios depend substantially on the epoch. Alternatively a time-dependent factor (see next section) can be introduced that changes the influence of an explanatory variable according to the epoch; significance of this factor relative to a constant effect would point to non-proportionality. Finally, we note that if non-proportional hazards are suspected for a factor that is not of primary interest, most software will allow this to be specified as a 'stratification factor', which means that separate underlying hazard functions are allowed for each category of that factor.

The term *relative risk* is often used interchangeably both with *hazard ratio* and with *odds ratio* (derived from logistic regression), so perhaps the term is best avoided. Hazard and odds ratios will be different for the same data set, since the odds ratios relate to a particular time while hazard ratios are concerned with the whole survival experience.

The comments in Section 7.3 concerning the dangers of automatic variable selection in logistic regression apply equally to Cox regression.

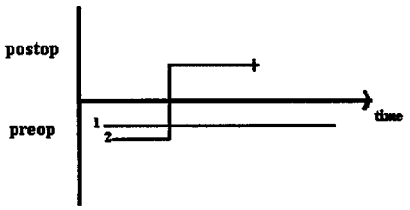


Figure 7. Patient 2 transfers from pre-operative to post-operative risk set at time of first operation

9. SURVIVAL WITH ONE TIME-DEPENDENT FACTOR

(Warning to readers: if earlier sections seemed difficult, perhaps now is the time to turn to the discussion! Section 11)

9.1. Factors which change with time

In randomized studies the intervention of interest is assumed to occur at the point of randomization, and hence the treatment groups are distinguished from the beginning of follow-up. By contrast, in observational studies we may encounter interventions that occur at any point in the period of follow-up, and yet we may be interested in making statements about the effectiveness of the intervention (relative to similar patients who have not had the intervention) in prolonging the time until a specified adverse event.

The risk associated with an intervention can only be assessed subsequent to the intervention – we may therefore consider intervention as a *time-dependent factor*, one for which a patient may change categories over time. In our example, we shall treat the factor *operation* as time-dependent with categories 'pre-op' and 'post-op'. All individuals start off as part of the 'pre-op' risk-set but some move to the 'post-op' risk set at the time of their operation (Figure 7).

Table X incorporates these transitions, showing the number of individuals in each risk set at the time of each death; the table provides the basis for comparing patients of the same elapsed time since presentation who are in the pre- and post-operative risk set. For example, from Figure 1, patient 24 shifts from the pre-op to the post-op risk set between the deaths of patients 16 and 11. In this way, we are comparing at any point the risk of those with the same survival time, who have had and did not yet have an operation. The layout of the table follows those shown previously and we may calculate estimated hazards and provide logrank test statistics as in Tables VI and VII.

The hazard ratio in the first year since presentation is estimated at 0·74/0·23 = 3·2, while from ages 1 to 10 the estimate is 3·5, suggesting that the proportional hazards assumption is realistic.

9.2. An error to avoid

In wishing to assess the influence of operation, it could be tempting to make a direct comparison between the post-presentation survival of patients who did and did not have an operation. If this approach were followed, with patients divided from entry into a 'no operation' and an 'operative' group, the results would have been somewhat different (Table XI).

The hazard in a table prepared this way would look better for those who had an operation, since their risk sets (the hazard denominator) are systematically inflated by including the pre-operative course of those who were later to have an operation. In this instance, the hazard ratio associated with operation is infancy is wrongly estimated to be 0·413/0·428 = 0·96, rather than the 3·2 obtained from the appropriate analysis in which patients switch risk sets.

Table X. Generation of pre-op and post-op hazard estimates

Question:	influence of operation on survival from presentation	
Analysis specification:		
inclusion criteria	all patients	
outcome	for pre-operative group:	preopded = 1
	for post-operative group:	dead = 1
time origin	presentation	
entry time	(a) presentation	0
	(b) at operation	opfpres
censoring rule	withdrawn at end of study	
survival time	for pre-operative group:	
	time from presentation until first op or to when last seen	unopfpre
	for post-operative group:	
	duration of follow-up	followup
period of observation	for pre-operative group:	
	entry: at presentation	0
	withdrawal: at operation or when last seen alive	unopfpre
	for post-operative group:	
	entry: at operation	opfpres
	withdrawal: when last seen alive	followup
explanatory variables (fixed)	none	
explanatory variables (time-dependent)	operation	hadop, changing from 0 to 1 at opfpres

Output:

Patient	Event time	(a) pre-operative				(b) post-operative			
		at risk	observed	estimated hazard/year	estimated SE(hazard/year)	at risk	observed	estimated hazard/year	estimated SE(hazard/year)
to 1 year post presentation									
5	4	29	1			1	0		
15	14	26	1			3	0		
21	77	20	0			8	1		
19	88	19	0	1/29 + 1/26		8	1	1/8 + 1/8	
3	117	16	1	+1/16 + 1/11	0·12	8	0	+ 1/8 + 1/8	0·30
2	121	14	0	= **0·23**		8	1	1/9 + 1/8	
30	142	13	0			8	1	= **0·74**	
29	193	11	0			9	1		
14	247	11	0			8	1		
17	275	11	1			7	0		
1–10 years post presentation									
12	393	10	0			7	1		
28	1100	7	0			8	1	(1/7 + 1/8	
16	1791	5	1	(1/5)/9		7	0	+1/8 + 1/5)/9	0·03
24	2982	2	0	= **0·02**	0·02	8	1	= **0·07**	
11	3098	2	0			5	1		

Table XI. Impact of operation on survival from presentation estimated incorrectly by division into operated and not operated groups (infancy only)

Question:	pre-operative and post-operative survival	
Analysis specification:		
inclusion criteria	all patients	
outcome	death	death = 1
time origin	presentation	
entry time	presentation	0
censoring rule	withdrawn at end of study	
survival time	time from presentation until death or censored	followup
period of observation	from presentation until death or censored	0 to followup
	censored	
explanatory variables	operation	hadop

Output:

Patient	Event time	hadop = 0				hadop = 1			
		at risk	events	hazard/year	estimated SE(hazard/year)	at risk	events	hazard/year	estimated SE(hazard/year)
5	4	12	1			18	0		
15	14	11	1	1/12 + 1/11		18	0	1/18 + 1/17	
21	77	11	0	+ 1/9		18	1	1/15 + 1/14	
19	88	11	0	+ 1/7		17	1	1/13 + 1/12	
3	117	9	1	= **0·428**	0·048	17	0	= **0·413**	0·029
2	121	9	0			15	1		
30	142	9	0			14	1		
29	193	9	0			13	1		
14	247	9	0			12	1		
17	275	7	1			12	0		

The analysis in Table XI may appear obviously incorrect, but early studies of the benefits of heart transplantation took the time of being placed on the waiting list as the time origin and compared the survival from that origin of those who did and did not receive a transplant. Transplantation was shown to be beneficial (as we would expect since a major reason for not obtaining a transplant is early death while on the waiting list), but the errors in this method of evaluation were rapidly made clear in a classic paper.[27]

9.3. Caveats

There is, of course, a great danger in trying to draw inferences about the effectiveness of interventions from non-randomized studies, since patients have been *selected* for the treatment, though the issues surrounding selection are often not clear-cut. In order to have any confidence in the conclusions of such an analysis, we need to understand the main factors that might have influenced the choice of intervention (age, anatomy etc.) and to have explicitly controlled for them in the model; Moses[6] has recently encouraged explicit recording of the *reasons* for intervention. Thus, in order to relate subsequent changes in risk to the intervention, we would need to feel that if two patients were identical in terms of the factors included in the model at the time of intervention, the clinician's decision to intervene on one rather than the other might just as well

have been based on the toss of a coin – we shall term this an assumption of a *non-informative intervention.* Naturally, this is an ideal target, but indicates the need properly to account for severity measures. These may, however, be allowed to change with time using the methodology in this section.

This analysis has not made any attempt to account for factors determining the decision to operate.

10. COMPLEX SURVIVAL ANALYSIS

10.1. Fixed and time dependent factors using Cox modelling with late entry

Time-dependent factors can be examined in a Cox proportional hazards model, which also gives the opportunity to adjust for fixed factors and hence attempt to make the assumption of a non-informative intervention more tenable. Our final example illustrates a means of exploring the influence of operation on the 'natural history' of the disease, adjusting for a fixed risk factor. This brings together many issues demonstrated individually in the preceding sections, including that of late entry, since the time origin is now shifted back to birth and we are estimating age-specific risks. In this example, the effect of operation is 'turned on' at the time of operation, and, in addition, the post-operative phase is divided into three stages: the first 30 day period; the period between one and 6 months, and after 6 months post-operative. The model also incorporates the values of the fixed factor describing pulmonary artery anatomy (paanat), so that inferences about the hazard related to operation (or avoiding operation) can be made 'independent' of the pulmonary artery anatomy.

Table XII shows that, allowing for pulmonary artery anatomy, in the month after operation, the risk of mortality is estimated to be 12 times that of patients of the same age but not operated on. This excess risk decreases dramatically for those who survive one month, but even for those who survive six months there is still a suggestion of continuing increased risk.

10.2. Computation

From a purely technical point of view, this type of analysis requires careful attention in definition of factors and in ensuring the computer programs work correctly. There is a huge increase in the time required for computation when time-dependent covariates are included. The p-values for individual levels come from comparing estimated coefficients divided by their standard errors to standard normal tables.

10.3. Caveats

Aside from the technical problems of such an analysis, great care is required in the interpretation of the output. It is tempting to think of such analyses as obviating the need for randomized trials, since they appear to provide a means of evaluating therapeutic interventions from observational databases, while suitably adjusting for the effect of selection of cases through additional fixed and time-dependent covariates. (See Franklin *et al.*[28,29] for examples of such analyses.)

However, the plausibility of the assumption of a non-informative intervention must always be open to doubt, since it is unlikely one could ever fully control for the clinician's decision to intervene at one time rather than another. Nevertheless, it is possible to imagine a situation where similar patients might be reasonably randomized to immediate or delayed operation. With genuinely similar patients in the pre- and post-operative risk set, the kind of analysis described in

Table XII. Results of Cox model for effect of operation and pulmonary artery anatomy

Question:	influence of operation and pa anatomy on pattern of survival from birth	
Analysis specification:		
inclusion criteria	all patients	
outcome	death	dead = 1
time origin	birth	
entry time	presentation	agepres
censoring rule	withdrawn at end of study	
survival time	age last seen alive	agelast
period of observation	presentation until survival time	agepres to agelast
explanatory variables (fixed)	pa anatomy	paanat
(time-dependent)	operation	hadop, changing from 0 to 1 at ageopl 1 to 2 at ageopl + 30 2 to 3 at ageopl + 180

Output:

Factor	Category	Hazard ratio relative to baseline	95% CI		*p*-value
operation	pre-op	1·00			
	up to 1 month post-op	12·00	1·56	92·69	0·017
	1 to 6 months post-op	1·94	0·28	13·29	0·50
	> 6 months post-op	1·43	0·28	7·02	0·66
paanat	0	1·00			
	1	1·48	0·41	5·28	0·55

this section could then supply an understanding of the role of operation which is difficult to provide with an observational study.

11. DISCUSSION

There is a very reasonable determination to maximize the value and range of inferences that can be drawn from large databases. However, it is clear that even modest inferences can only be drawn at the cost of some assumptions; these assumptions are best made explicit and ideally should be tested. Some aspects, such as the independence of the censoring mechanism, will always be untestable however large the data set, and hence there will always be some reliance on background knowledge and clinical insight.

In contrast to the value placed on the conclusions of a randomized trial, the value placed on the conclusions of an observational study will depend largely on whether all potentially relevant factors have been examined. The onus is on the designers of the observational study to make these factors explicit and ensure that they are adequately represented in the data set. Given such representation, the statistical methods demonstrated here provide some tools for meaningful inter-centre comparison for audit, identification of explanatory variables, prediction on individual cases and so on. We emphasize, however, the range of more sophisticated statistical methods that are becoming available, for example in dealing with recurrent events[30] or adjustment of predictive models for over-fitting to a database.[31]

The proliferation of databases in many branches of medicine and surgery has been partly in the expectation that they would provide a way of examining some management issues which have seemed intractable to the randomized trial approach – whether because numbers of similar patients adequate to support a trial are not available or because an ethical trial is hard to design. We have argued that inferences about how good or bad an investment is afforded by an intervention is particularly difficult to assess when simply observing the outcome of even large numbers of patients. It is here that the intellectual basis of the randomized trial is most potent. However, the careful analysis of databases might crystallize a management problem which is amenable to a randomized trial – perhaps of non-conventional design – for example, randomizing patients to alternative timing of operation. One of the most valuable products of good databases should be the increased potential to design incisive and efficient confirmatory experiments.

APPENDIX I: SPSS COMMANDS FOR ANALYSES

Provides derived variables

```
COMPUTE followup = agelast – agepres.
COMPUTE opfpres = – 1.
IF (ageop1 > 0) opfpres = ageop1 – agepres.
COMPUTE unopage = agelast.
IF (ageop1 > 0) unopage = ageop1.
COMPUTE unopfpre = ageop1 – agepres.
IF (MISSING (ageop1)) unopfpre = followup.
COMPUTE preopded = 0.
IF (MISSING (ageop1) & dead = 1) preopded = 1.
COMPUTE hadop = 1.
IF (MISSING (ageop1)) hadop = 0.
RECODE
  agepres
  (Lowest thru 365 = 0) (366 thru Highest = 1) INTO agepresx.
COMPUTE dedlyrpp = 0.
IF (dead = 1 & followup = 365) dedlyrpp = 1.
IF (adfol = 0) dedlyrpp = 2.
```

Section 3.2 and Table II. Proportion dying within one year of presentation

```
COMPUTE filter_$ = (adfol = 1).
FILTER BY filter_$.
FREQUENCIES
  VARIABLES = dedlyrpp.
FILTER OFF.
```

Section 4.5, Table III and Figure 3. Non-parametric survival from presentation. Corresponding Weibull survival prepared using EGRET

```
KM
  followup  /STATUS = dead(1)  /PRINT TABLE  /PLOT SURVIVAL.
```

Section 4.7. Table IV and Figure 4. Survival function assuming all in risk set from birth. Survival function with late entry generated in EGRET

```
KM
  agelast  /STATUS = dead(1)  /PRINT TABLE  /PLOT SURVIVAL.
```

Section 5.2 and Table V outcome after 1 year. One factor at a time.

```
COMPUTE filter_$ = (adfol = 1).
FILTER BY filter_$.
CROSSTABS
  /TABLES = paanat agepresx BY dedlyrpp
  /FORMAT = AVALUE NOINDEX BOX LABELS TABLES
  /CELLS = COUNT ROW.
```

Section 7.1 and Table V. Outcome after 1 year. More than 1 explanatory variable.

```
LOGISTIC REGRESSION dedlyrpp
  /METHOD = ENTER paanat agepresx
FILTER OFF.
EXECUTE.
```

Section 6.1 and Figure 6. Survival with one fixed explanatory variable

```
KM
  followup  /STRATA = paanat  /STATUS = dead(1)
  /PRINT TABLE
  /PLOT SURVIVAL.
```

Section 8.1 and Table IX. Cox regression using whole survival experience. First one variable at a time, then together.

```
COXREG
  folloup  /STATUS = dead(1)
  /METHOD = ENTER paanat
  /PRINT = CI (95)
  /CRITERIA = ITERATE (20).
COXREG
  followup  /STATUS = dead(1)
  /METHOD = ENTER sex
  /PRINT = CI (95)
  /CRITERIA = ITERATE (20).
COXREG
  followup  /STATUS = dead(1)
  /METHOD = ENTER paanat sex
  /PRINT = CI (95)
  /CRITERIA = ITERATE (20).
```

Section 9.1 and Table X, estimating hazard ratio associated with operation: in this SPSS analysis this ratio is assumed constant over whole period after presentation.

First create a logical time-dependent covariate T_COV_ that is 1 if the patient had an operation AND time-since-presentation > interval-to-operation.

```
TIME PROGRAM.
  COMPUTE T_COV_ = (hadop = 1) & (T_ > opfpres).
  Fit time-dependent covariate

COXREG
  followup  /STATUS = dead(1)
  /METHOD = ENTER T_COV_
  /ITERATE(20).
```

Analysis in 10.1, Cox with time dependent factors and late entry performed using EGRET.

APPENDIX II: A NON-TECHNICAL GLOSSARY OF TERMS

Adjusted odds ratio: the odds ratio for one explanatory variable assuming other explanatory variables in the model remain fixed. Derived by logistic regression.
Baseline hazard function: the hazard function for a patient in the baseline category of all the variables entered into, say, a Cox regression analysis.
Baseline odds: the odds on the outcome of interest occurring for a patient in the baseline category of all the variables entered into a logistic regression analysis.
Censoring: withdrawal from the study before the event of interest has occurred, because the study has ended without this event occurring or for other reasons specified in the study design.
Cox regression: this technique deals with outcomes occurring over the whole survival experience and allows the generation of adjusted hazard ratios for multiple factors to be estimated simultaneously; it requires a proportional hazards assumption.
Entry time: the time when a patient starts contributing to the study. In randomized studies or observational studies where all patients have come under observation before the study starts (for example, studies of survival after surgery) the entry time and time origin of the study will be identical. However, for some observational studies, the patient may not start follow-up until after the time origin of the study and these patients contribute to the study group only after their 'late entry'.
Explanatory variables (*also risk factors, covariates, predictors, independent variables*): quantities which may be associated with better or worse outcome.
Factor: an explanatory variable with a limited number of states, possibly a continuous variable which has been divided up into discrete categories.
Hazard function: the instantaneous risk of a patient experiencing a particular event at each specified time.
Hazard ratio: the hazard associated with one category of patient divided by the hazard associated with another category. The hazard ratio can be estimated at an instant or averaged over an interval.

Informative censoring: when withdrawal from the study may not be independent of the current hazard; if patients at higher or lower risk than the rest are withdrawn, this will introduce bias.
Informative late entry: when time of entry is itself a predictor of survival time, perhaps because it reflects severity of the condition concerned additional to that expressed by measured risk factors.
Late entry (*left truncation*): this occurs when patients come under observation after the time origin of the study. In terms of their survival outlook, these patients may or may not be the same as those already in the risk set.
Logistic regression: this technique deals with prediction of outcome at a fixed time interval after the time origin and allows adjusted odds ratios for multiple factors to be estimated simultaneously; it assumes such odds ratios are independent of underlying risk and (unless interaction terms are fitted) the values other factors take on.
Non-informative intervention: an assumption that each patient who underwent an intervention did so for a reason which was not related to their underlying risk or, if it were related, that this relationship can be understood in terms of other associated variables entered into the analysis.
Non-parametric survival function: an estimate of the survival function that depends only on the size of the risk set at the time each event occurs, and hence the graph proceeds by downward steps.
Odds ratio (*unadjusted, simple, univariate odds ratio*): the odds associated with one category of patient divided by the odds associated with a 'baseline' category of patient.
Odds: a measure of risk defined as $p/(1-p)$, where p is the probability of the event in question.
Outcomes (*events, responses or dependent variables*): the endpoint of interest (outcomes dealt with in this paper have all been configured as binary events).
Parametric survival function: an assumption that the survival function is governed by a small number of parameters which are estimated from the data; the graph of the parametric survival function is smooth.
Period of observation: interval between the entry time and the occurrence of the event or censoring.
Proportional hazards: this important assumption is fulfilled if two categories of patient are being compared and their hazard ratio is constant over time (though the instantaneous hazards may vary).
Relative risk: this term can confuse as it sometimes is taken to mean a hazard ratio ('relative risk' over the whole survival experience) and sometimes an odds ratio ('relative risk' over a fixed time interval).
Risk set: the set of patients in the study at a specified time.
Semi-parametric: 'parametric' assumptions may be made about some aspects of a model, while other components may be estimated 'non-parametrically'. In the Cox regression procedure, a parametric model for the relative hazard is overlaid on a non-parametric estimate of baseline hazard.
Survival function: the probability of being free of the event at a specified time.
Survival time: interval between the time origin and the occurrence of the event or censoring.
Time-dependent factor: sometimes factors which come into play after the time origin of the study require consideration because of their possible influence on the probability of the subsequent occurrence of an adverse outcome. To compare the outcomes of patients who have had and who have not yet had this event, two risk sets are compared; patients transfer from one risk set to the other at the time of occurrence of the event of interest.
Time origin: the beginning of the story the study aims at telling. In observational studies, the patients may come under observation before or after the time origin of the study.

ACKNOWLEDGEMENT

Kate Bull is supported by the British Heart Foundation.

REFERENCES

1. Kirlin, J. W., Blackstone, E. H., Tchervenkov, C. I., Casteneda, A. R. and the Congenital Heart Surgeons Society. 'Clinical outcomes after the arterial switch operation for transposition: patient, support, procedural and institutional risk factors', *Circulation*, **86**, 1501–1515 (1992).
2. Hanley, F. L., Sade, R. M., Blackstone, E. H., Kirlin, J. W., Freedom, R. M. and Nanda, N. C., 'Outcomes in neonatal pulmonary atresia and intact ventricular septum', *Journal of Thoracic and Cardiovascular Surgery*, **105**, 406–427 (1993).
3. Byar, D. P., Simon, R. M., Friedewald, W. T. et al. 'Randomised clinical trials: perspectives on some recent ideas', *New England Journal of Medicine*, **295**, 74–80 (1976).
4. Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J. and Smith, P. G. 'Design and analysis of randomised clinical trials requiring prolonged observation of each patient. Introduction and design', *British Journal of Cancer* **34**, 585–612 (1976).
5. Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J. and Smith, P. G. 'Design and analysis of randomised cinical trials requiring prolonged observation of each patient. Analysis and examples', *British Journal of Cancer*, **35**, 1–39 (1976).
6. Moses, L. E., 'Measuring effects without randomized trials? Options, problems, challenges', *Medical Care* **33**, AS8–14 (1995).
7. D'Agostino, R. B. and Kwan, H. 'Measuring effectiveness. What to expect without a randomized control group', *Medical Care*, **33**, AS95–105 (1995).
8. Healy, M. J. R. 'Survival data', *Archives of Disease in Childhood*, **73**, 374–377 (1995).
9. Altman, D. G. *Practical Statistics for Medical Research*, Chapman and Hall, London 1991.
10. Clayton, D. and Hills, M. *Statistical Models in Epidemiology*, Oxford University Press, Oxford, 1993.
11. Fisher, L. D. and van Belle, G. *Biostatistics: a Methodology for the Health Sciences*, Wiley, New York, 1993.
12. Cox, D. R. and Oakes, D. *Analysis of Survival Data*, Chapman and Hall, London, 1984.
13. Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. *Statistical Models Based on Counting Processes*, Springer, New York, 1993.
14. D'Agostino, R. B., Lee, M-L., Belanger, A. J., Cupples, L. A., Anderson, K. and Kannel, W. B. 'Relation of pooled logistic regression to time-dependent Cox regression analysis: the Framingham Heart Study', *Statistics in Medicine*, **9**, 1501–1516 (1990).
15. Sackett, D. L., Haynes, R. B. and Tugwell, P. *Clinical Epidemiology: a Basic Science for Clinical Medicine*, 2nd edn., Little, Brown, Boston, 1991.
16. Dambrosia, J. M. and Ellenberg, J. H. 'Statistical considerations for a medical data base', *Biometrics*, **36**, 323–332 (1980).
17. Bull, C., Somerville, J., Ty, E. and Spiegelhalter, D. J. 'Presentation and attrition in complex pulmonary atresia', *Journal of the American College of Cardiology*, **25**, 491–499 (1995).
18. Fleiss, J. L. *Statistical Methods for Rates and Prorportions*, 2nd edn, Wiley, New York, 1981.
19. Cnaan, A. and Ryan, L. 'Survival analysis in natural history studies of disease', *Statistics in Medicine*, **8**, 1255–1268 (1989).
20. Keiding, N. 'Independent delayed entry (with discussion)', in Wein, J. P. and Goel, P. K. (eds), *Survival Analysis: State of the Art*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1992, pp. 309–328.
21. Blackstone, E. H., Naftel, D. C., Turner, M. E. 'The decomposition of time varying hazard into phases, each incorporating a separate stream of concomitant information', *Journal of the American Statistical Association*, **81**, 615–624 (1986).
22. Tsai, W. Y. 'Testing the assumption of independence of truncation time and failure time', *Biometrika*, **77**, 169–177 (1990).
23. Rowan, K. M., Kerr, J. H., Major, E., McPherson, K., Short, A. and Vessey, M. P. 'Intensive Care Society's APACHE II study in Britain and Ireland-II: Outcome comparisons of intensive care units after adjustment for case mix by the American APACHE II method', *British Medical Journal*, **307**, 977–981 (1993).

24. Harrell, F. E., Lee, K. L. and Mark, D. B. 'Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors', *Statistics in Medicine*, **5**, 361–388 (1996).
25. MacMohan, S., Peto, R., Cutler, J., Collins, R., Sorlie, P., Neaton, J., Abbott, R., Godwin, J., Dyer, A. and Stamler, J. 'Blood pressure, stroke and coronary heart disease. Part 1. Prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias', *Lancet*, **355**, 765–774 (1990).
26. Hess, K. R. 'Graphical methods for assessing violations of the proportional hazards assumption in Cox regression', *Statistics in Medicine*, **14**, 1707–1724 (1995).
27. Mantel, N. and Byar, D. P. 'Evaluation of response-time data involving transient states: an illustration using heart transplant data', *Journal of the American Statistical Association*, **69**, 81–86 (1974).
28. Franklin, R. C. G., Spiegelhalter, D. J., Anderson, R. H., Macartney, F., Rossi-Filho, R. I., Douglas, J. M., Rigby, M. L. and Deanfield, J. E. 'Double inlet ventricle presenting in infancy. iii: Outcome and potential for definitive repair', *Journal of Thoracic and Cardiovascular Surgery*, **101**, 924–934 (1991).
29. Franklin, R. C. G., Spiegelhalter, D. J., Sullivan, I. D., Anderson, R. H., Thoele, D. S., Shinebourne, E. A. and Deanfield, J. E. 'Tricuspid atresia presenting in infancy: survival and suitability for the Fontan operation', *Circulation*, **87**, 427–439 (1993).
30. Clayton, D. G. 'Some approaches to the analysis of recurrent event data', *Statistical Methods on Medicial Research*, **3**, 244–262 (1994).
31. van Houwelingen, H. C. and Thorogood, J. 'Construction, validation and updating of a prognostic model for kidney graft survival', *Statistics in Medicine*, **14**, 1999–2008 (1995).

TUTORIAL IN BIOSTATISTICS
METHODS FOR INTERVAL-CENSORED DATA

JANE C. LINDSEY* AND LOUISE M. RYAN

Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A.

SUMMARY

In standard time-to-event or survival analysis, occurrence times of the event of interest are observed exactly or are right-censored, meaning that it is only known that the event occurred after the last observation time. There are numerous methods available for estimating the survival curve and for testing and estimation of the effects of covariates in this context. In some situations, however, the times of the events of interest may only be known to have occurred within an interval of time. In clinical trials, for example, patients are often seen at pre-scheduled visits but the event of interest may occur in between visits. These data are interval-censored. Owing to the lack of well-known statistical methodology and available software, a common *ad hoc* approach is to assume that the event occurred at the end (or beginning or midpoint) of each interval, and then apply methods for standard time-to-event data. However, this approach can lead to invalid inferences, and in particular will tend to underestimate the standard errors of the estimated parameters. The purpose of this tutorial is to illustrate and compare available methods which correctly treat the data as being interval-censored. It is not meant to be a full review of all existing methods, but only those which are available in standard statistical software, or which can be easily programmed. All approaches will be illustrated on two data sets and compared with methods which ignore the interval-censored nature of the data. We hope this tutorial will allow those familiar with the application of standard survival analysis techniques the option of applying appropriate methods when presented with interval-censored data.

Statist. Med., **17**, 219–238 (1998)

1. INTRODUCTION

In a standard survival analysis application, individuals are followed over time for the occurrence of a specific event. If the event is observed to occur, the data is recorded as the time the event occurred, T, and the censoring indicator δ, is set to the value 1. If by the end of the period of observation the event has not been observed, the observation is considered to be right-censored, the value of T would be set to the last observation time and δ would take the value 0. If a patient

* Correspondence to: Jane C. Lindsey, Department of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, U.S.A. E-mail: lindsey@sdac.harvard.edu

Contract grant sponsor: National Institute of Allergy and Infectious Diseases
Contract grant number: 1u01 AI 41110

Contract grant sponsor: National Cancer Institute
Contract grant number: CA 48061

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
 ISBN: 0-470-02365-1

was recruited to a trial and the event of interest had already occurred, their data would be left-censored, and T would be set to 0. If data are either observed exactly or are right-censored, the most common situation in most clinical trial settings, there are numerous parametric, semi-parametric and non-parametric methods available for estimation of the survival curves, and for hypothesis testing and estimation of the effects of covariates of interest. Most are readily applicable using standard statistical software (SAS,[1] S-plus,[2] GLIM,[3] Stata[4] etc.).

In some situations, however, the times of the events of interest may only be known to have occurred within an interval of time, $[L, R]$, where $L \leqslant T \leqslant R$. This can occur in a clinical trial, for example, when patients are assessed only at pre-scheduled visits. If the event has not occurred at one visit (at time L) but has by the following visit (at time R), T is known only to within the interval $[L, R]$. These are known as interval-censored data. Note that exactly observed, right- and left-censored data are special cases of interval-censored data, with $L = R$ for exactly observed data, $R = \infty$ for right-censored and $L = 0$ for left-censored observations. As with the right-censored case, a data analyst would be interested in estimation of the survival curves and in assessing the importance of covariates. Since few statistical packages explicitly accommodate these kind of data, a common approach is to assume that the event occurred at the end (or beginning or midpoint) of each interval, and then apply methods for standard time-to-event data. As discussed by Rücker and Messerer,[5] Odell *et al.*[6] and Dorey *et al.*[7] however, this can lead to biased and misleading results.

Although few methods for analysing interval-censored data are widely used, parametric models are available in SAS[1] and S-plus.[2] In addition, some specialized methods for estimation of the survival curve and for hypothesis testing can be easily programmed in standard statistical packages. We will illustrate selected methods using two data sets, given in Tables I and II.

The first[8] is from a retrospective study of patients with breast cancer designed to compare radiation therapy alone versus in combination with chemotherapy with respect to the time to cosmetic deterioration. This data set has been analysed by several authors to illustrate various methods for interval censored data. Patients were seen initially every 4 to 6 months, with decreasing frequency over time. If deterioration was seen, it was known only to have occurred between two visits. Deterioration was not observed in all patients during the course of the trial, so some data were right-censored. The second example[9] focuses on the development of drug resistance (measured using a plaque reduction assay[10]) to zidovudine in patients enrolled in four clinical trials for the treatment of AIDS. Samples were collected on the patients at a subset of the scheduled visit times dictated by the four protocols. Since the resistance assays were very expensive, there were few assessments on each patient, resulting in very wide intervals, $[L, R]$, if resistance was seen to have occurred, and a high proportion of right-censored observations. Because of the sparseness of these data, this is a challenging data set to analyse. Of interest were the effects of stage of disease, dose of zidovudine and CD4 lymphocyte counts at time of randomization on the time to development of resistance.

In this tutorial, our aim is to illustrate how these kinds of data can be analysed using existing methodology either with software already available in standard statistical packages, or by writing simple code. We describe parametric and non-parametric methods for estimation of the survival distribution and for testing the effect of covariates in the following section. The methods are illustrated using the two data sets described above in Section 3 followed by some conclusions and recommendations.

Table I. Breast cancer data set: a value of Right = 61 implies data are right-censored

Therapy = 1		Therapy = 0	
Left	Right	Left	Right
8	12	45	61
0	22	6	10
24	31	0	7
17	27	46	61
17	23	46	61
24	30	7	16
16	24	17	61
13	61	7	14
11	13	37	44
16	20	0	8
18	25	4	11
17	26	15	61
32	61	11	15
23	61	22	61
44	48	46	61
14	17	46	61
0	5	25	37
5	8	46	61
12	20	26	40
11	61	46	61
33	40	27	34
31	61	36	44
13	39	46	61
19	32	36	48
34	61	37	61
13	61	40	61
16	24	17	25
35	61	46	61
15	22	11	18
11	17	38	61
22	32	5	12
10	35	37	61
30	34	0	5
13	61	18	61
10	17	24	61
8	21	36	61
4	9	5	11
11	61	19	35
14	19	17	25
4	8	24	61
34	61	32	61
30	36	33	61
18	24	19	26
16	60	37	61
35	39	34	61
21	61	36	61
11	20		
48	61		

Table II. AIDS data set: a value of Right = 26 implies data are right-censored

Left	Right	Stage	Dose	CD4 100–399	CD4 ⩾ 400
0	16	0	0	0	1
15	26	0	0	0	1
12	26	0	0	0	1
17	26	0	0	0	1
13	26	0	0	0	1
0	24	0	0	1	0
6	26	0	1	1	0
0	15	0	1	1	0
14	26	0	1	1	0
12	26	0	1	1	0
13	26	0	1	0	1
12	26	0	1	1	0
12	26	0	1	1	0
0	18	0	1	0	1
0	14	0	1	0	1
0	17	0	1	1	0
0	15	0	1	1	0
3	26	1	0	0	1
4	26	1	0	0	1
1	11	1	0	0	1
13	19	1	0	0	1
0	6	1	0	0	1
0	11	1	1	0	0
6	26	1	1	0	0
0	6	1	1	0	0
2	12	1	1	0	0
1	17	1	1	1	0
0	14	1	1	0	0
0	25	1	1	0	1
2	11	1	1	0	0
0	14	1	1	0	0

2. DESCRIPTION OF METHODS

2.1. Parametric Methods

The most straightforward approach, which can be implemented directly in SAS[1] (and the *survreg* routine in S-plus,[2] although we will not go into details for this here), is to assume a parametric model for the failure times. The SAS[1] procedure LIFEREG provides a way of fitting accelerated failure time (AFT) models for a variety of distributions to interval censored data. The AFT model is defined by the transformation

$$T_{\mathbf{z}} = T_0 e^{-\beta \mathbf{z}}$$

where $T_{\mathbf{z}}$ is the failure time random variable for an individual with covariate $\mathbf{z}$ and T_0 is the failure time the individual would have if they had covariate value 0. The effect of changing covariates is to shrink or stretch the time to event. If β is negative, then the covariate has the effect of 'speeding up time' so that individuals with larger values of $\mathbf{z}$ have higher failure rates and hence

shorter survival times. The survival function can be written as

$$S(t; \mathbf{z}) = P(T_{\mathbf{z}} \geqslant t \mid \mathbf{z}) = P(T_0 \geqslant te^{\beta \mathbf{z}}) = S_0(te^{\beta \mathbf{z}})$$

where $S_0(t)$ is the survivorship function for an individual with covariate value 0. Taking natural logarithms, the AFT model can be expressed as

$$\log T = \log T_0 - \boldsymbol{\beta}\mathbf{z}.$$

If we assume that $\log T_0$ can be expressed as $\mu + \sigma W$, where W is a random variable, then the model can be written in a linear model-like form:

$$\log T = \mu - \boldsymbol{\beta}\mathbf{z} + \sigma W. \tag{1}$$

The PROC LIFEREG module of SAS[1] fits this model, except that the sign is changed on the regression coefficients. That is, SAS fits

$$\log T = \mu + \boldsymbol{\beta}\mathbf{z} + \sigma W. \tag{2}$$

SAS allows a variety of distributions to be placed on the error term W, including the log of the exponential, log-normal and log-gamma distributions. The intercept parameter μ and the scale parameter σ are usually not of direct interest, although for some distributions, there is a relationship between the AFT model and a proportional hazards model through the scale parameter. For example, if W is an extreme value distribution (log of a unit exponential), then T has a Weibull distribution. Note that because of the change in sign implicit in the AFT formulation, the direction of covariate effects will be opposite to those fit with a Cox proportional hazards model. This is discussed further in Section 3. Hypothesis testing is straightforward through the use of standard likelihood theory. The AFT framework provides a flexible and simple way to conduct a wide variety of analyses. An example of the SAS code required to fit such a model is given in Section 3.

2.2. Piecewise Exponential Model

Another parametric model which can be useful is the piecewise exponential model. Suppose we break the time scale into J intervals $I_j = (\tau_{j-1}, \tau_j]$ for $j = 1, \ldots, J$, and assume a constant hazard in interval j,

$$\lambda(t) = \lambda_j \quad \text{for } t \in I_j.$$

Covariate effects can be accommodated using proportional hazards. That is, if $\mathbf{z}$ represents a vector of covariates and ψ the corresponding regression coefficients, we write

$$\lambda(t \mid \mathbf{z}) = \lambda_j e^{\psi \mathbf{z}}.$$

The piecewise exponential model has the advantage that as J increases, the model becomes more non-parametric in nature. This allows for flexibility in modelling, while still providing the advantages of a parametric method in terms of hypothesis testing and estimation. Although no standard statistical packages allow fitting of the piecewise exponential model explicitly, Farrington[11] describes a macro for GLIM,[3] which uses theory based on generalized linear models. Alternatively, a method using the expectation maximization (EM) algorithm[12] is described in the Appendices. This approach can be used to write specialized software in any computer pacakge. As before, hypothesis testing can be done using standard likelihood theory, as long as the number of intervals does not become too large.

2.3. Non-parametric estimation of survival curve

Peto[13] was the first to propose a non-parametric method for estimating the survival distribution based on interval-censored data. Turnbull[14] derived the same estimator, but used a different approach in estimation. Suppose T_i, $(i = 1, \ldots, n)$, the survival times for n patients, are independent random variables with left continuous survival function $S(t) = \Pr(T \geqslant t)$. If the T_i are not observed directly, but instead are known to lie in the interval $[L_i, R_i]$, then the likelihood for the n observations is

$$L = \prod_{i}^{n} \{S(L_i) - S(R_i^+)\}. \tag{3}$$

By $S(t^+)$, we mean

$$\lim_{\Delta \to 0^+} S(t + \Delta)$$

which may be different from $S(t)$, since $S(t)$ is left continuous. (See Cox and Oakes[15] for a more complete explanation.) It is important to note here that different authors vary in their conventions regarding definition of the censoring interval. We follow the convention of Peto[13] and Turnbull,[14] who assume a closed interval, $[L_i, R_i]$. This definition facilitates the accommodation of observations that are known exactly, that is, $L_i = R_i$, but necessitates the use of the R_i^+ notation in equation (3) to allow a non-zero contribution to the likelihood for these observations. Finkelstein[16] assumes semi-closed censoring intervals, so needs to add the convention that the likelihood contribution for any observation with an exact failure time, T_i, is $S(t_i)$. Good arguments can be made for and against almost any convention for defining the censoring intervals. In practice, the choice will have little impact and any reasonable convention can be adopted.

With right-censored data, Kaplan and Meier[17] showed that the closed form product limit estimator is the generalized maximum likelihood estimate. This curve jumps at each observed event time. With interval-censored data, although there is no closed-form solution, the generalized maximum likelihood estimator can be shown to take a similar form, but with jumps on a discrete set of 'equivalence classes' defined through the intervals. This will be discussed in more detail presently. Peto[13] described an algorithm to find the maximum likelihood estimates using a programmed search which was practical with relatively small data sets but too computationally intense to be used in general. Turnbull[14] derived the same estimator using an iterative self-consistency algorithm, described below. Gentleman and Geyer[18] show that this self-consistent estimator is not always the maximum likelihood estimator (MLE), and that the MLE is not necessarily unique and discuss conditions under which this can be determined. One problem with applying their methodology is that there is no publicly available software.

Since the observed event times are known to occur only within potentially overlapping intervals, the survival curve can only jump within so-called equivalence sets $[q_j, p_j], j = 1, \ldots, m$, where $q_j \leqslant p_j < q_{j+1} \leqslant \cdots$. Between p_j and q_{j+1} the curve is flat. The estimate of $S(t)$ is unique only up to these equivalence classes, that is, any function that jumps the appropriate amount within the equivalence class will yield the same likelihood. A simple algorithm to identify these equivalence classes goes as follows. First, string out in increasing order, all the timepoints $\{L_i\}$ and $\{R_i\}$ as shown in Figure 1. Then attach labels L and R to indicate which times correspond to left or right hand censoring limits, respectively. Equivalence classes are then defined by the regions between each left hand limit that is immediately followed by a right hand limit. In the example depicted in Figure 1, there are three left/right pairs and a final left-censored time.

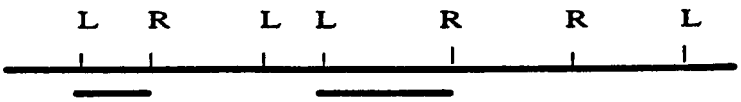


Figure 1. Example of equivalence classes for interval-censored data

However, there are only two equivalence classes within which the estimated survival curve will jump. The equivalence classes for the AIDS[9] data set, derived using the above algorithm are [6, 6], [12, 12], [14, 14], [15, 15], [17, 17].

Once the equivalence classes have been determined, Turnbull's[14] self-consistency algorithm is derived as follows. Let A be an n by m matrix of indicators with the i,jth value set to 1 if $[L_i, R_i]$ covers $[q_j, p_j]$ and 0 otherwise. This corresponds to allowing a positive probability for the ith event to occur in equivalence class j if the two overlap. Also define P_0 to be an m by 1 vector of starting probabilities corresponding to the jump sizes between q_j and $p_j, j = 1, \ldots, m$. This can be set to $1/m$. The following SAS[1] IML code will find the self-consistent estimator:

```
Q = 1;
EPS = 0·0001;
DO UNTIL (ABS (Q − P) < EPS);
  Q = P;
  C = A*P;
  C = C + (C = 0)# EPS;
  P = P# (A'*(1/C))/N;
  END;
```

N is the number of observations in the data set and EPS is the tolerance for absolute change in parameter values from one iteration to the next.

Although no standard statistical packages produce the survival curve estimate, it is easy to implement in any computer programming language. To ensure the algorithm finds a correct solution, starting values must place mass in each equivalence class. Peto[13] and Turnbull[14] suggest using the inverse of the second derivatives of the observed data log-likelihood to obtain estimates of the standard error for the survival curve. Second derivatives may be found numerically or by applying the method of Louis.[19] Neither approach is particularly straightforward, however. Furthermore, there is as yet, no theoretical justification for this procedure in the literature, the problem being a violation of the usual assumption of a fixed, unknown parameter that remains unchanged with increasing sample size.

2.4. Logspline estimation of the survival curve

Kooperberg and Stone[20] provide software (logspline.fit, available through Statlib for S-plus[2]) which can be used to obtain smoothed estimates of the survival function based on interval-censored data using splines. We will not go into detail about their methodology, which is

described fully in their paper. Briefly though, they fit smooth functions to the log-density function of the failure times within subsets of the time axis defined by the 'knots', and constrained to be continuous at those points. This provides a loosely parametric framework for finding estimates of the survival and hazard functions which can be useful for exploratory data analysis. Although the approach is likelihood based, their software at this time does not allow for hypothesis testing. Their approach is related to that of Rosenberg,[20] who uses splines to model the hazard function, but software for Rosenberg's[21] method is not publicly available.

2.5. Non-parametric hypothesis testing

We introduce one non-parametric method for hypothesis testing that allows comparison of two groups with left-, right- or interval-censored data developed by Finkelstein.[16] Other methods have been introduced in the literature, but few are used because of the lack of easily available software.

For right-censored data, the two sample logrank test statistic can be written as

$$U = \sum_{j=1}^{m} \left(d_{1j} - \frac{d_j n_{1j}}{n_j} \right)$$

where m is the number of event times, d_{1j} is the number of events in the first group and d_j the number of events in both groups at event time j. Similarly, n_{1j} is the number at risk of an event in the first group, and n_j the number at risk in both groups at event time j. The statistic U, divided by its standard error, can be compared to a standard normal distribution to test for differences in survival times between the two groups.

For interval-censored data, Finkelstein[16] derives a score test, U', which looks like the standard logrank test, but uses 'pseudo' counts of events and numbers at risk, rather than the observed counts in the standard logrank test. The pseudo counts are the expected numbers of events and individuals at risk at each jump in the survival curve q_j, p_j, calculated by summing the probabilities that each individual failed and the probability that they were at risk at q_j. Her test has the form:

$$U' = \sum_{j=1}^{m} \left(d'_{1j} - \frac{d'_j n'_{1j}}{n'_j} \right).$$

Now m is the number of equivalence classes $[q_j, p_j]$, d'_{1j} is the 'pseudo' number of events in the first group in that interval and d'_j is the 'pseudo' number of events in both groups in that interval. Similarly, n'_{1j} is the 'pseudo' number at risk of an event in the first group, and n'_j is the 'pseudo' number at risk in both groups. The statistic U', divided by its standard error, can be compared to a standard normal distribution. This statistic is easily calculated, and a FORTRAN[22] program to perform the analysis for one categorical covariate can be obtained from the author. In addition, So[23] provides a macro to find the self-consistent estimates using maximization routines in PROC IML.[24] His method allows for multiple categorical covariates, but since the algorithm depends on numerical routines, it can become unstable if m is small or the number of covariates is large.

3. EXAMPLES

In this section, we apply the methods described in the previous section to the breast cancer and AIDS data sets. Results are compared with the naive approaches that make assump-

tions about when the events took place, for example at the beginning, midpoint or end of each interval.

The breast cancer data set (Table I), described more fully in Finkelstein and Wolfe,[8] consists of 94 observations from a retrospective study looking at the time to cosmetic deterioration. Information is available on one covariate, the type of therapy – either radiation alone (coded 0), or in combination with chemotherapy (coded 1). Of the 94 observations, 56 are interval-censored and 38 are right-censored.

The AIDS data[9] are compiled from patients at one study site participating in four trials in the AIDS Clinical Trials Group. Since assays to assess drug resistance are expensive, relatively few data are available. Of the 31 patients with at least one assay, 13 are right-censored (their last measurement showed no resistance to zidovudine). Since all patients are assumed to be sensitive to drug at randomization, 18 are interval-censored (one measurement showed sensitivity and the next showed resistance (5 individuals) or only one measurement was available after randomization which was resistant (13 individuals)). Covariate information is available on stage of disease (early (0) or late (1)), dose of zidovudine (low (0) or high (1)) and CD4 lymphocyte count at baseline (< 100, 100–399, and $\geqslant 400$ cells/mm^3 at randomization (coded with two indicator variables using the < 100 group as baseline).

3.1. Estimation of time to event

We compare five ways of estimating the time to event, ignoring the effects of covariates. First assume exact times of event are known and use a standard Kaplan–Meier[17] estimator. This can be done by either assuming the event occurred at the left interval, or at the right interval. These two extreme cases should roughly bracket the estimates from the interval-censoring methods. A second approach is to correctly model the interval-censored nature of the data using the method proposed by Turnbull.[14] Third, using a Weibull model, the survivorship function can be modelled using the estimates from SAS PROC LIFEREG.[1] Next, using the Statlib software in S-plus,[2] the spline models of Kooperberg and Stone[20] are applied to each data set. Finally, we estimate the survivorship function from the piecewise model using

$$\hat{S}(t) = \prod_{j=1}^{J-1} e^{-\lambda_j(\tau_j - \tau_{j-1})} e^{-\lambda_j(t-\tau_{j-1})}.$$

Kaplan–Meier estimates of the survival curve can be found using SAS PROC LIFETEST.[1] Two columns of data are required: one giving the failure time T_i (FTIME) and a second giving the censoring indicator (EVENT). The computer code is:

```
PROC LIFETEST;
  TIME FTIME * EVENT (0);
  TITLE 'KAPLAN MEIER ESTIMATE OF SURVIVAL CURVE';
```

Note that in the above SAS code, EVENT(0) indicates that a value of 0 in the EVENT variable means the value was right-censored.

The Weibull survival curve can be found using SAS PROC LIFEREG.[1] Two columns of data are required: the left endpoint (LEFT) and the right endpoint (RIGHT). If the data are right-censored, then RIGHT is set to a missing value ("." in SAS). If the data are left-censored,

LEFT is set to a missing value. The computer code is:

```
PROC LIFEREG;

  MODEL (LEFT, RIGHT) = /D = WEIBULL;

  TITLE 'WEIBULL FAILURE TIME SURVIVAL CURVE';
```

Once the S-plus[2] routines have been loaded, the commands for fitting the spline estimates are straightforward. Three variables are needed: the left (LEFT) and right (RIGHT) intervals, and a censoring (STATUS) variable. The function *logspline. fit* is used to fit the model and *logspline.plot* to plot the results:

```
right[status == 0] ← left[status == 0]

fit ← logspline.fit(right = right[status == 0],

                    interval = cbind(left[status == 1], right[status == 1]))

logspline.plot(fit = fit, n = 1000, what = "s")
```

Where:

```
   right = collects all right-censored observations
interval = collects all interval-censored observations
     fit = fit uses results from logspline.fit output
       n = 1000 asks for 1000 equally spaced points to be plotted
    what = "s" request survival curve
         = "h" requests hazard function
```

Note that left-censored and exactly observed data can also be included.

The Turnbull[14] estimate is found using the algorithm presented in the previous section. The piecewise exponential models are fit using the approach given in the Appendices using a SAS[1] macro.

Estimates are shown in Figures 2 and 3 for the breast cancer and resistance data sets, respectively. For the breast cancer example, the Kaplan–Meier estimates, as expected, bracket the Turnbull estimate. The piecewise exponential, fit here with intervals at 10, 15, 20, 25, 30, 35, 40 and 61 months, tracks the Turnbull curve and lies very close to the estimate from the Weibull, as does the logspline estimate.

The estimated survival curve for the AIDS data takes very few steps in the non-parametric models, reflecting the high degree of censoring in this small data set. The Kaplan–Meier estimates no longer bracket the Turnbull estimate, mainly because the Turnbull estimate has very few jumps due to the particular configuration of this data set. The Weibull estimate gives a very similar estimate to the piecewise model. The logspline estimate also tracks the parametric models closely, although it is smoother than the piecewise exponential model. For these data, the non-parametric methods are not very helpful in understanding the data.

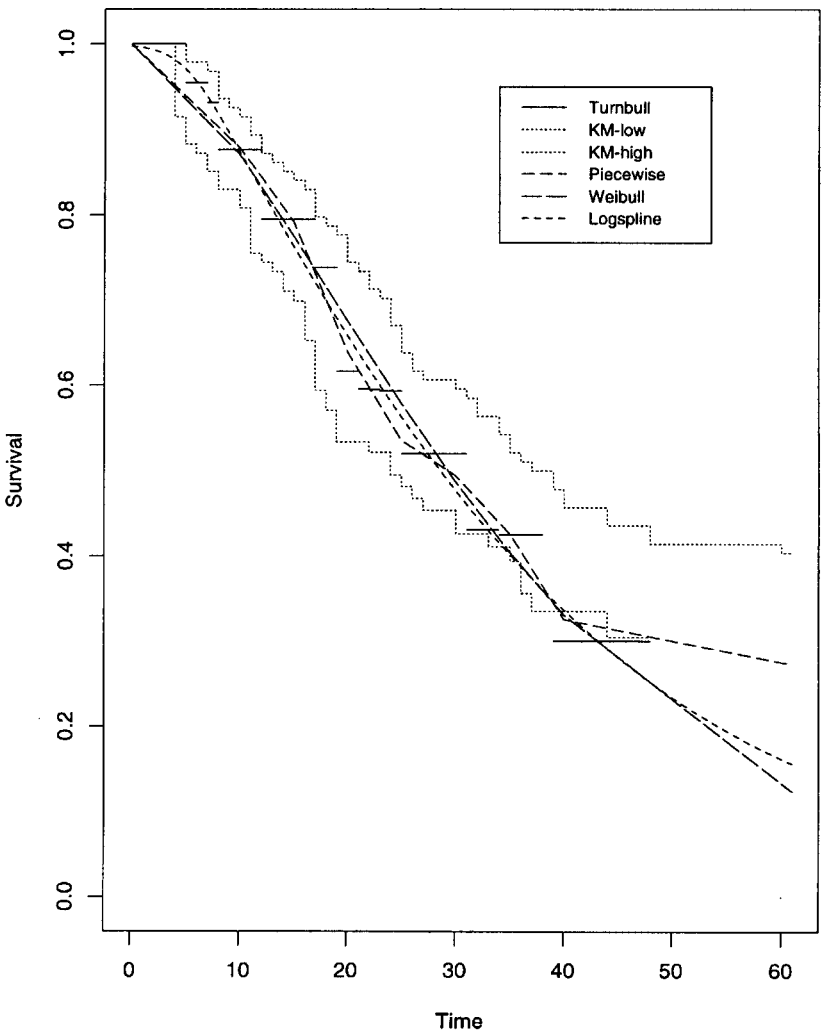


Figure 2. Survival curve estimates for breast cancer data with no covariates

3.2. Covariate effects on time to event

We initially consider single covariate models. Results using the Finkelstein[16] model which correctly accounts for the interval-censored nature of the data, the Cox[25] regression models which assume exact event times (taken to be the left, midpoint and right extremes of the interval), the piecewise model and the exponential, Weibull and log-normal models are shown in Tables III and IV for the breast cancer and resistance data, respectively.

Finkelstein's[16] methodology is implemented using So's[23] algorithm, since the FORTRAN program available from Finkelstein[16] does not give parameter estimates or standard errors for the covariate term. Standard errors are estimated using a numerical approximation to the second

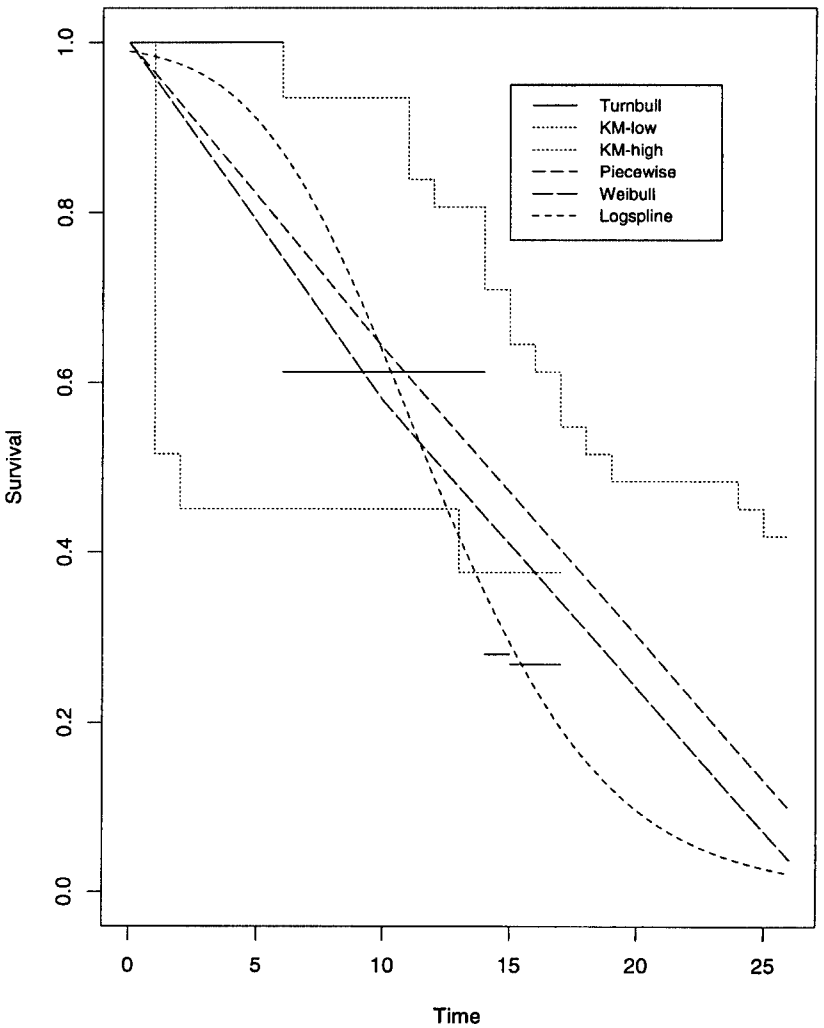


Figure 3. Survival curve estimates for AIDS data with no covariates

derivative matrix using the NLPFDD[24] subroutine. As stated earlier, however, we are not fully justified in using these standard errors which rely on as yet unproven asymptotic theory. In practice, however, these standard error estimates are usually acceptable.

To fit Cox[25] regression models, failure times (FTIME) and censoring indicators (EVENT) are input to the SAS PROC PHREG,[1] with a list of covariates, which for the breast cancer data is just THERAPY, coded 0 or 1:

```
PROC PHREG;
  TIME FTIME*EVENT(0) = THERAPY;
  TITLE 'COX REGRESSION';
```

Table III. Breast cancer data: effect of therapy on time to event

Model	Estimate	Standard error	*p*-value
Finkelstein	0·791	0·288	0·005
Cox (left)	0·912	0·287	0·001
Cox (midpoint)	0·900	0·285	0·001
Cox (right)	0·768	0·286	0·006
Exponential	− 0·742	0·277	0·006
Weibull (therapy)	− 0·568	0·176	< 0·001
(scale)	0·619	0·074	
Lognormal (therapy)	− 0·421	0·203	0·037
(scale)	0·882	0·097	
Piecewise (8 intervals)	0·930	0·287	< 0·001

Table IV. AIDS data: effect of stage of disease and dose of zidovudine on time to event

Model	Estimate	Standard error	*p*-value
Stage of disease			
Finkelstein	.	.	.
Cox (left)	0·766	0·490	0·112
Cox (midpoint)	1·454	0·496	0·003
Cox (right)	0·837	0·502	0·092
Exponential	− 1·308	0·514	0·010
Weibull	− 0·719	0·271	0·002
(scale)	0·393	0·139	
Log-normal	− 0·847	0·241	< 0·001
(scale)	0·388	0·130	
Piecewise	1·727	0·584	0·003
Dose of zidovudine			
Finkelstein	0·948	0·744	0·166
Cox (left)	0·527	0·533	0·307
Cox (midpoint)	0·698	0·578	0·202
Cox (right)	0·497	0·543	0·346
Exponential	− 0·584	0·543	0·266
Weibull	− 0·353	0·285	0·189
(scale)	0·490	0·189	
Log-normal	− 0·392	0·375	0·271
(scale)	0·718	0·256	
Piecewise	0·712	0·560	0·188

To fit parametric models with covariates, using an exponential distribution as an example:

```
PROC LIFEREG;
  MODEL (LEFT, RIGHT) = THERAPY/D = EXPONENTIAL;
  TITLE 'EXPONENTIAL FAILURE TIME MODEL';
```

Estimates from Finkelstein's[16] model, Cox[25] regression and the piecewise models are directly comparable because they model multiplicative changes in the baseline hazard. Positive parameter estimates indicate higher failure rates for individuals with larger values of the covariate. The exponential model parameter should be of comparable magnitude to the Cox[25] and Finkelstein[16] models, but with the sign reversed. The reason for the reversed sign, as explained in Section 2.1, is that PROC LIFEREG[1] and most exponential regression packages report the estimated parameters from the accelerated failure time model (1). The Weibull is the only family of models that is both proportional hazards and AFT. It can easily be shown that the estimated regression coefficient should be comparable to the coefficients from the Cox[25] and Finkelstein[16] models, after being divided by its scale parameter. For the breast cancer data example, $0{\cdot}568/0{\cdot}619 = 0{\cdot}918$, which is the same order of magnitude as the Finkelstein[16] and Cox[25] models. A similar relationship holds for the log-normal model, although less concordance is expected since it is not a proportional hazards model. All reported p-values are from likelihood ratio tests.

For the breast cancer data, the four analyses give similar results qualitatively. All show an increased hazard for the group on radiation and chemotherapy which is statistically significant. Note the minimal impact of the differing assumptions about timing of events on the Cox analysis. In this data set, treating the data as exact time versus interval-censored makes little difference.

For the AIDS data, we look at the individual effects of stage and dose on the time to development of resistance. For stage, all methods indicate an increased risk of developing resistance for the patients in a later stage of disease. The estimates from the Finkelstein[16] model are so unstable they are not shown in the table, reflecting that for these data, the non-parametric methods do not perform as well as the parametric methods. Unlike the breast cancer data, changing assumptions about when events are assumed to occur has a big impact on the Cox analysis. Assuming the event occurred at the left interval effectively decreases the sample size by 13 patients, resulting in a non-significant result. The strength of the significance is also affected when the midpoint or right extreme of the interval is used as the exact event time. With such large effects, using a method which accounts for the interval-censored nature of the data is preferable, but with so few steps in the survival curve using the non-parametric methods, a parametric analysis is the best choice. Similar trends are seen in fitting the effect of dose. The Cox model results are highly dependent on the assumptions about when the event occurred. No methods show a significant effect of dose on the time to development of resistance, although the trend for a quicker time to resistance for the higher dose group is an interesting observation, as clinicians had previously believed that patients on lower doses of ZDV might have become resistant faster than those on a higher dose.

For the AIDS data, we can do a more extensive analysis using the piecewise constant hazards model (see Table V). CD4 lymphocyte count, coded as an indicator variable, is an important predictor alone, but provides no additional predictive power when included in a model with stage of disease ($p = 0{\cdot}542$). This is not surprising as CD4 lymphocyte count will be highly correlated with stage of disease.

One of the decisions needed in an analysis based on the piecewise method, is the number and breakpoints of the intervals for the underlying hazard. In the standard survival analysis context, Friedman[26] recommends choosing intervals so that the expected number of events in each interval is comparable. Choosing two intervals with the breakpoint of 10 months gives approximately 10 and 8 expected number of events in the two intervals in a model with no covariates. Additional results for univariate models are shown in Table V for two intervals with a break at

Table V. Analysis of AIDS data with piecewise constant hazards model and two intervals: individual covariate models

Breakpoint	Effect	Estimate	Standard error	p-value
10	Stage	1·727	0·584	0·003
	Dose	0·712	0·560	0·188
	Baseline CD4: 100–399	−1·384	0·724	0·015
	: ⩾ 400	−1·872	0·764	
	Stage and Baseline CD4			0·542
5	Stage	1·770	0·599	0·002
	Dose	0·628	0·548	0·237
	Baseline CD4: 100–399	−1·407	0·623	0·015
	: ⩾ 400	−1·813	0·654	
5, 10	Stage	1·771	0·548	0·002
	Dose	0·714	0·567	0·190
	Baseline CD4: 100–399	−1·408	0·658	0·014
	: ⩾ 400	−1·870	0·698	

5 months and three intervals with breaks at 5 and 10 months. The break at 5 months has an unbalanced expected number of events at the maximum likelihood estimates (4 and 14) and takes much longer to converge using the EM algorithm than with a single break at 10 months. However, conclusions for the three covariates are the same regardless of the choice of interval. Stage of disease and CD4 lymphocyte are important predictors while dose is not. More advanced patients develop resistance faster. Patients with higher CD4 lymphocyte counts have a lower risk of developing resistance. Alternatively, as described by Rosenberg,[21] one can choose a large number of breakpoints and use penalized likelihood techniques to smooth the estimates and avoid numerical problems associated with overparameterization. Caution would be needed in interpreting the asymptotic results if the number of intervals approached the number of events. Although further work needs to be done in choosing the best number of intervals, it seems based on the results presented here that the method is fairly robust to the choice of interval.

4. CONCLUSIONS

Interval-censored data often occur in medical applications. Although only two of the major statistical packages (SAS[1] and S-plus[2]) have procedures for analysing these data using parametric models, some non-parametric methods are easily programmed and their use should be considered. In particular, Turnbull's[14] method for non-parametric estimation of the survival distribution, Kooperburg and Stone's[20] logspline estimates of the survival function and Finkelstein's[16] test for covariates are recommended. As seen in the AIDS data set, when data are heavily censored, making assumptions about when events occurred and using techniques such as Cox regression can lead to inaccurate conclusions. It can also result in unstable estimation in the non-parametric methods. The parametric methods available in SAS[1] and S-plus[2] are the most readily available alternative. As seen with the examples presented in this paper, these parametric

approaches can be highly satisfactory in their performance. This is especially so if one chooses the Weibull or log-normal family that allows a reasonably wide range of distributional shapes. To allow more flexible modelling with weaker parametric assumptions, we suggest the use of a piecewise constant hazards model. This approach can be programmed using an EM algorithm or using a macro available in GLIM[11] and is useful for estimation of the survival curve as well as hypothesis testing.

Other methods have been proposed in the literature for testing, but none seem to be used routinely, most likely due to the lack of availability of software. Finkelstein and Wolfe,[8] Self and Grossman,[27] Miller[28] and Buckley and James[29] have all proposed tests for assessing the covariate effects. Although Borgan *et al.*,[30] Chiang *et al.*[31] and Brookmeyer and Goedert[32] all describe more complicated problems, they discuss the issue of interval censoring as a special case. All propose using piecewise-constant hazards to model at least part of multi-state disease processes. Note that piecewise exponential models can be thought of as a special case of the method proposed by Rosenberg[21] since they correspond to modelling the hazards with 0-order splines.

APPENDIX I: EM ALGORITHM FOR PIECEWISE EXPONENTIAL MODEL

The log-likelihood for the piecewise exponential model is a complicated formula involving sums of integrals which have no closed form solutions. While a variety of numerical maximization methods are available, the use of an EM algorithm[12] turns out to provide a numerically stable approach to finding the maximum likelihood estimates. It consists of two steps: the maximization (M-step) of a 'complete data likelihood', assuming the sufficient statistics are known, and finding expected values of the sufficient statistics of the complete data likelihood given the observed data and the current parameter estimates (E-step). The approach works well when the complete data likelihood is easily maximized. The algorithm iterates between the two steps until convergence.

The first task in setting up the EM algorithm is to define the 'complete' data. One choice of complete data could be the exact failure times. However, it turns out that fitting the piecewise exponential model is also straightforward when there is right-censoring. Hence in our context, we need to define as the complete data, only the failure times of the individuals experiencing an event and the censoring times of the right-censored observations. A simple likelihood can be formed as follows. Suppose we break the time scale into J intervals $I_j = (\tau_{j-1}, \tau_j]$ for $j = 1, \ldots, J$, and assume a constant hazard in interval j

$$\lambda(t) = \lambda_j \text{ for } t \in I_j.$$

If T_i, the exact times of right-censoring ($\delta_i = 0$) or failure ($\delta_i = 1$), are known for $i = 1, \ldots, n$, the log-likelihood from a standard piecewise exponential model can be simply expressed as a function of the numbers of events and times at risk within each interval. Consider contributions to the complete data likelihood. If the time t of death or censoring falls in the jth interval, the likelihood contribution (ignoring covariates) is:

$$q(t) = \lambda(t)^{\delta_i} e^{-\int_0^t \lambda(u)\,du} = \lambda_j^{\delta_i} \exp\left\{ -\left[\sum_{k=0}^{j-1} \lambda_k(\tau_k - \tau_{k-1}) \right] - \lambda_j(t - \tau_j) \right\}.$$

Taking logs and summing over all individuals, it is straightforward to show that the complete data log-likelihood can be written as:

$$L_\lambda = \sum_{j=1}^{J} \{N_j \ln \lambda_j - \lambda_j S_j\}$$

where N_j is the number of subjects experiencing an event in interval j and S_j is the 'person-time at risk' for an event in interval j. This is also proportional to the likelihood obtained by assuming the N_j are Poisson distributed with mean parameter $\lambda_j S_j$. With no covariates, the MLEs of the hazards have the closed form solution:

$$\hat{\lambda}_j = \frac{N_j}{S_j} \tag{4}$$

for $j = 1, \ldots, J$. With covariates, an iterative algorithm as described in Laird and Olivier[33] can be used to find the MLEs. Alternatively, a Poisson regression package can be used treating the outcome as N_j and $\log S_j$ as an offset.

The E-step of the EM algorithm finds the expectation of the complete data sufficient statistics, conditional on the observed data ($\boldsymbol{Y}$) and substituting the current parameter estimates $\hat{\boldsymbol{\lambda}}$. The sufficient statistics to be estimated are the number of subjects experiencing an event in each interval, N_j and the time at risk, $S_j, j = 1 \ldots J$. If person i is right-censored ($\delta_i = 0$), then they do not contribute to N_j, and their contribution to the time at risk in interval j is known exactly. Suppose person i fails ($\delta_i = 1$) in $V_i = [L_i, R_i]$. Their contribution to N_j is the conditional probability, $p_j(V_i)$, that their event occurred in interval j, given that $t_i \in V_i$. Summing over all individuals who failed, then

$$E(N_j | \boldsymbol{Y}, \hat{\boldsymbol{\lambda}}) = \sum_{\{i:\delta_i = 1\}} p_j(V_i).$$

The form of $p_j(V_i)$ is given in Appendix II.

The conditional expected time at risk in the jth interval is

$$E(S_j | \boldsymbol{Y}, \hat{\boldsymbol{\lambda}}) = \sum_{i:t_i \in I_j \& \delta_i = 0\}} (t_i - \tau_{j-1}) + \sum_{\{i:t_i \in I_l \& \delta_i = 0, l > j\}} (\tau_j - \tau_{j-1}) + \sum_{\{i:\delta_i = 1\}} E_j(V_i).$$

The first term is the contribution of the patients who are right-censored ($\delta_i = 0$) in interval j. The second term captures the time contributed by the patients who are censored at some time after the jth interval ($L_i > \tau_j$). The last term accounts for the contribution of the interval-censored individuals who have an event in V_i to the jth interval. The term $E_j(V_i)$ represents the expected time at risk in the jth interval for someone who fails in V_i. Detailed expressions are given in Appendix II.

The first iteration of the EM algorithm requires starting values for the parameter estimates. One approach is to begin with some sensible assumptions about when a patient has an event within an interval. For example, one could assume that the patient experienced the event half way through the interval. Times at risk and numbers of events can be calculated based on this assumption and values for the underlying hazards estimated using equation (4). Starting values for the covariate effects can be set to zero. Estimates of standard errors for the parameters can be calculated using the methods of Meng and Rubin.[34] Tests of hypothesis concerning the effects of these covariates can be made, either using a likelihood ratio test, Wald or score test.

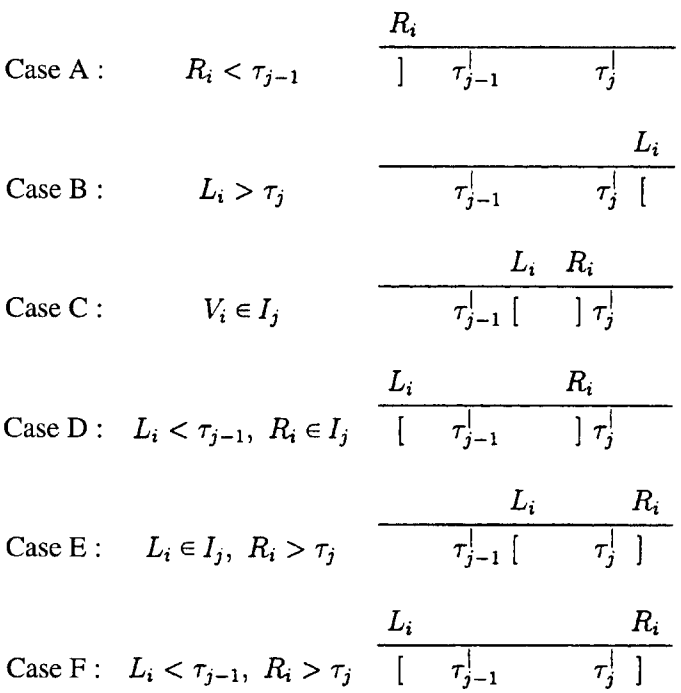
APPENDIX II: PROBABILITIES AND EXPECTED TIMES AT RISK FOR INDIVIDUALS WHO FAIL FOR PIECEWISE EXPONENTIAL MODEL

First define:

$$V_i = \text{interval of event for individual } i = [L_i, R_i]$$

$$q(t) = \text{probability of event at } t = \lambda(t)^{\delta_i} e^{-\int_0^t \lambda(u)\,du}.$$

There are six possible cases, determined by the location of V_i with respect to I_j:



Consider first $p_j(V_i)$, the conditional probability of failing in interval I_j, given the individual's failure interval V_i:

$$p_j(V_i) = \frac{1}{\int_{L_i}^{R_i} q(t)\mathrm{d}t} \begin{cases} 0 & \text{Case A or B} \\ \int_{L_i}^{R_i} q(t)\mathrm{d}t & \text{Case C} \\ \int_{\tau_{j-1}}^{R_i} q(t)\mathrm{d}t & \text{Case D} \\ \int_{L_i}^{\tau_j} q(t)\mathrm{d}t & \text{Case E} \\ \int_{\tau_{j-1}}^{\tau_j} q(t)\mathrm{d}t & \text{Case F.} \end{cases}$$

The $p_j(V_i)$ are relatively straightforward. In case C, for example, V_i is contained in interval I_j, so the probability that the event occurred in interval j is 1. The expected times at risk are slightly

more complicated.

$$E_j(V_i) = \begin{cases} 0 & \text{Case A} \\ (\tau_j - \tau_{j-1}) & \text{Case B} \\ \dfrac{\int_{L_i}^{R_i} tq(t)\,dt}{\int_{L_i}^{R_i} q(t)\,dt} - L_i & \text{Case C} \\ p_j(V_i)\left\{\dfrac{\int_{\tau_{j-1}}^{R_i} tq(t)\,dt}{\int_{\tau_{j-1}}^{R_i} q(t)\,dt} - \tau_{j-1}\right\} & \text{Case D} \\ p_j(V_i)\left\{\dfrac{\int_{L_i}^{\tau_j} tq(t)\,dt}{\int_{L_i}^{\tau_j} q(t)\,dt} - L_i\right\} + \sum_{l>j} p_l(V_i)(\tau_j - \tau_{j-1}) & \text{Case E} \\ p_j(V_i)\left\{\dfrac{\int_{\tau_{j-1}}^{\tau_j} tq(t)\,dt}{\int_{\tau_{j-1}}^{\tau_j} q(t)\,dt} - \tau_{j-1}\right\} + \sum_{l>j} p_l(V_i)(\tau_j - \tau_{j-1}) & \text{Case F.} \end{cases}$$

Cases A to D are straightforward for the expected times at risk, but we describe case F in more detail. The expected time contributed to I_j for each individual is 0 times the probability they failed before τ_{j-1} ($\sum_{l<j} p_l(V_i)$), plus $p_j(V_i)$ times a proportion of the interval $(\tau_j - \tau_{j-1})$ (the first term in the case F equation), plus the complete width of I_j times the probability of failing after τ_j ($\sum_{l>j} p_l(V_i)$). This latter term is the second part of the equation in case F. Case E can be described using similar logic.

ACKNOWLEDGEMENTS

This work was supported by the Center for Biostatistics in AIDS Research of the Pediatric AIDS Clinical Trials Group, under the National Institute of Allergy and Infectious Diseases contract No. 1u01 AI 41110, and grant number CA48061 from the National Cancer Institute.

REFERENCES

1. *SAS/STAT Users Guide*, SAS Institute, Inc., North Carolina, 1990.
2. *SPlus User's Manual*, Statistical Sciences Inc., Seattle, 1991.
3. Francis, B., Green, M. and Payne, C. *The GLIM System: Release 4 Manual*, Clarendon Press, Oxford, 1993.
4. *Stata Reference Manual*, Stata Press, Texas, 1995.
5. Rücker, G. and Messerer, D. 'Remission duration: an example of interval-censored observations', *Statistics in Medicine*, **7**, 1139–1145 (1988).
6. Odell, P. M., Anderson, K. M. and D'Agostino, R. B. 'Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model', *Biometrics*, **48**, 951–959 (1992).
7. Dorey, F. J., Little, R. J. and Schenker, N. 'Multiple imputation for threshold-crossing data with interval censoring', *Statistics in Medicine*, **12**, 1589–1603 (1993).
8. Finkelstein, D. M. and Wolfe, R. A. 'A semi-parametric model for regression analysis of interval censored failure time data', *Biometrics*, **41**, 933–945 (1985).
9. Richman, D. D., Grimes, J. M. and Lagakos, S. W. 'Effect of stage of disease and drug dose on zidovudine susceptibilities of isolates of human immunodeficiency virus', *Journal of AIDS*, **3**, 743–746 (1990).
10. Larder, B. A., Darby, G. and Richman, D. D. 'HIV with reduced sensitivity to zidovudine isolated during prolonged therapy', *Science*, **243**, 1731–1734 (1989).
11. Farrington, C. P. 'Interval censored survival data: a generalized linear modelling approach', *Statistics in Medicine*, **15**, 283–292 (1996).

12. Dempster, A. P., Laird, N. M. and Rubin, D. B. 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of The Royal Statistical Society, Series B*, **39**, 1–38 (1977).
13. Peto, R. 'Experimental survival curves for interval-censored data', *Applied Statistics*, **22**, 86–91 (1973).
14. Turnbull, B. W. 'The empirical distribution function with arbitrarily grouped, censored and truncated data', *Journal of The Royal Statistical Society, Series B*, **38**, 290–295 (1976).
15. Cox, D. R. and Oakes, D. *Analysis of Survival Data*, Chapman and Hall University Printing House, Cambridge, U.K. 1984, p. 13.
16. Finkelstein, D. M. 'A proportional hazards model for interval-censored failure time data', *Biometrics*, **42**, 845–854 (1986).
17. Kaplan, E. L. and Meier, P. 'Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association*, **53**, 457–481 (1958).
18. Gentleman, R. and Geyer, C. J. 'Maximum likelihood for interval censored data: consistency and computation', *Biometrika*, **81**, 618–623 (1994).
19. Louis, T. 'Finding the observed information matrix when using the EM algorithm', *Journal of the Royal Statistical Society, Series B*, **44**, 226–233 (1982).
20. Kooperberg, C. and Stone, C. J. 'Logspline density estimation for censored data', *Journal of Comput. and Graph. Stat.*, **1**, 301–328 (1992).
21. Rosenberg, P. S. 'Hazard function estimation using B-splines', *Biometrics*, **51**, 874–887 (1995).
22. Miessner, L. P. and Organick, E. I. *FORTRAN 77*, Addison-Wesley, Reading, Massachusetts, 1984.
23. So, Ying, 'Interval-censored survival data', *Proceedings of the 19th Annual SAS Users Group*, SAS Institute, North Carolina, 1107–1113 (1994).
24. *SAS/IML Software*, SAS Institute, Inc., North Carolina, 1990.
25. Cox, D. R. 'Regression models and life tables (with discussion)', *Journal of the Royal Statistical Society, Series B*, **74**, 187–220 (1972).
26. Friedman, M. 'Piecewise exponential models for survival data with covariates', *Annals of Statistics*, **10**, 101–113 (1982).
27. Self, S. G. and Grossman, E. A. 'Linear rank tests for interval-censored data with application to PCB levels in adipose tissue of transformer repair workers', *Biometrics*, **42**, 521–530 (1986).
28. Miller, R. G. 'Least squares regression with censored data', *Biometrika*, **63**, 447–464 (1976).
29. Buckley, J. and James, I. 'Linear regression with censored data', *Biometrika*, **66**, 429–436 (1979).
30. Borgan, Ø., Liestol, K. and Ebbesen, P. 'Efficiencies of experimental designs for an illness-death model', *Biometrics*, **40**, 627–638 (1984).
31. Chiang, Y., Hardy, R. J., Hawkins, C. M. and Kapadia, A. S. 'An illness-death process with time-dependent covariates', *Biometrics*, **45**, 669–681 (1989).
32. Brookmeyer, R. and Goedert, J. J. 'Censoring in an epidemic with an application to Hemophilia-associated AIDS', *Biometrics*, **45**, 325–335 (1989).
33. Laird, N. and Olivier, D. 'Covariance analysis of censored survival data using log-linear analysis techniques', *Journal of the American Statistical Association*, **76**, 231–240 (1981).
34. Meng, X. and Rubin, D. 'Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm', *Journal of the American Statistical Association*, **86**, 899–909 (1991).

TUTORIAL IN BIOSTATISTICS

ANALYSIS OF BINARY OUTCOMES IN LONGITUDINAL STUDIES USING WEIGHTED ESTIMATING EQUATIONS AND DISCRETE-TIME SURVIVAL METHODS: PREVALENCE AND INCIDENCE OF SMOKING IN AN ADOLESCENT COHORT

JOHN B. CARLIN[1,*†], RORY WOLFE[1], CAROLYN COFFEY[2] AND GEORGE C. PATTON[2]

[1] *Clinical Epidemiology and Biostatistics Unit, Royal Children's Hospital Research Institute and University of Melbourne, Department of Paediatrics, Royal Children's Hospital, Parkville, Vic 3052, Australia*
[2] *Center for Adolescent Health, University of Melbourne, Parkville, Vic 3052, Australia*

SUMMARY

Longitudinal studies are increasingly popular in epidemiology. In this tutorial we provide a detailed review of methods used by us in the analysis of a longitudinal (multiwave or panel) study of adolescent health, focusing on smoking behaviour. This example is explored in detail with the principal aim of providing an introduction to the analysis of longitudinal binary data, at a level suited to statisticians familiar with logistic regression and survival analysis but not necessarily experienced in longitudinal analysis or estimating equation methods. We describe recent advances in statistical methodology that can play a practical role in applications and are available with standard software. Our approach emphasizes the importance of stating clear research questions, and for binary outcomes we suggest these are best organized around the key epidemiological concepts of prevalence and incidence. For *prevalence* questions, we show how unbiased estimating equations and information-sandwich variance estimates may be used to produce a valid and robust analysis, as long as sample size is reasonably large. We also show how the estimating equation approach readily extends to accommodate adjustments for missing data and complex survey design. A detailed discussion of gender-related differences over time in our smoking outcome is used to emphasize the need for great care in separating longitudinal from cross-sectional information. We show how *incidence* questions may be addressed using a discrete-time version of the proportional hazards regression model. This approach has the advantages of providing estimates of relative risks, being feasible with standard software, and also allowing robust information-sandwich variance estimates.

* Correspondence to: John B. Carlin, Clinical Epidemiology and Biostatistics Unit, Royal Children's Hospital Research Institute and University of Melbourne, Department of Paediatrics, Royal Children's Hospital, Parkville, Vic 3052, Australia. E-mail: j.carlin@medicine.unimelb.edu.au

† Current address (to December 1999): John Carlin, Department of Epidemiology & Biostatistics, College of Public Health, MDC-56, University of South Florida, 13201 Bruce B. Downs Blvd., Tampa, FL33612 USA.

Contract/grant sponsor: National Health and Medical Research Council, Australia

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
 ISBN: 0-470-02365-1

1. INTRODUCTION

In this tutorial we discuss and illustrate statistical approaches to the analysis of longitudinal studies in which measurements have been made repeatedly on a cohort of subjects at a sequence of time points. Studies collecting data of this kind are increasingly common in epidemiology, where the focus of the underlying research questions may be on developing an understanding of the causal sequence of various behaviour and disease events (observational studies) or on characterizing and comparing the course of disease in different treatment groups (clinical trials). In the social sciences, such studies are generally called panel studies.

The tutorial arises from the analysis of a longitudinal study of adolescent health in Victoria, Australia. Our aim here is to document in detail the methods of statistical analysis that have been used in substantive articles that concentrate on results.[1,2] We also aim to review the technical rationale for the methods used, at a level accessible to applied statisticians who do not require full mathematical details. Our choice of methods is largely motivated by the capabilities of current software packages and we provide detailed examples of code that was used to implement the analysis in Stata.[3]

There has been rapid development in methods for the statistical analysis of longitudinal data over the last decade or so. In the biostatistics literature, attention focussed initially on applications of the multivariate normal distribution for modelling continuous outcome measures.[4] More recently much work has been done on the more difficult problems of analysing binary[5,6] and ordinal[7] outcomes. Although this recent work has had great impact on strategies used in the practical analysis of real data problems, it is often difficult for beginners to penetrate the technicalities surrounding the various methods that have been proposed. The complexities increase rapidly when some of the common complications of real epidemiological data are considered, for example, missing data, complex survey design and measurement error. Furthermore, as more elaborate statistical techniques become available to tackle more complex questions, it is increasingly difficult for the applied researcher to keep up with these technical developments and to determine when the more complex approach is likely to pay off. Our tutorial aims to bring modern methods of longitudinal analysis for binary data within the grasp of statisticians and epidemiologists who have a sound grounding in logistic regression and proportional hazards (Cox) regression, but may have no experience in longitudinal analysis or the methods we discuss including estimating equations and robust estimation of standard errors.

In the next section we provide a brief overview of the adolescent health study from which our data arise, in particular outlining some of the study's principal research questions. The subsequent section describes how a first set of questions, which can be described as relating to *prevalence* of particular unhealthy behaviours, was addressed. These analyses used the so-called 'marginal' modelling approach, where no attempt is made to examine the interrelationship within the same individual between outcomes at different points in time. The meaning of the term 'marginal modelling' is explained, and later contrasted with the idea of 'subject-specific' modelling. We consider in some detail alternative methods available for estimating marginal association between a binary outcome and predictor variables, including logistic regression with 'robust' standard errors and generalized estimating equations, so our treatment serves as a simple introduction to the idea of estimating equations. Finally, the estimating equation method is used with incorporation of weights to take account of the stratified survey design and to provide a simple adjustment for missing data. The fourth section turns to more specifically longitudinal questions concerned with the *incidence* or uptake of behaviours. We detail a method of discrete

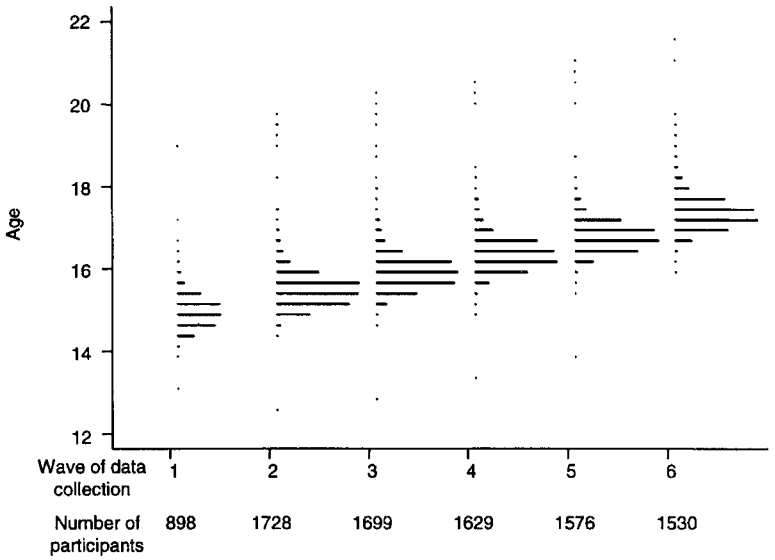


Figure 1. Age distribution of the participants at each wave of data collection

survival analysis that enables regression techniques to be applied for relating incidence of a behaviour, such as commencement of regular smoking, to subject-level covariates such as psychological well-being and peer smoking levels. Finally we discuss some of the major issues that have not been adequately addressed by our analyses to date.

2. THE DATA AND RESEARCH QUESTIONS

The Adolescent Health Cohort Study was a longitudinal study of teenagers in the state of Victoria carried out between August 1992 and July 1995. The basic design was to measure a cohort of approximately 2000 adolescents on six occasions or waves of data collection, at six-monthly intervals. Such studies are sometimes called 'multiwave' or 'panel' designs because the same 'panel' of participants is assessed at each wave of the study; another term commonly used is 'repeated measures' cohort study. A complication with our study was that about half the cohort was not recruited until the second wave, resulting in the pattern of data shown by the age frequency histograms of Figure 1.

The initial sample of participants was identified in a cross-sectional survey using a two-stage sampling procedure.[1] Schools were selected with probability proportional to the number of year 9 students (within 10 strata defined by geographic regions and type of school — classified as Government, Catholic or Independent), and within each school a single intact class was selected at random. At this initial wave, students had a mean age of 14·9 years; nearly all had passed their fourteenth birthday but few had reached their sixteenth (Figure 1). At the second wave of data collection, six months later, a second intact class was selected from each participating school. Subsequent waves of data were collected at six-monthly intervals over 3 years, resulting in an intended six timepoints for the primary cohort and five timepoints for the second sample. A total

sample of 2032 students (1003 recruited at wave 1 and 1029 at wave 2) was identified from 44 schools.

As in all cohort studies of this kind, there were problems of sample attrition and missing data. Of the total sample of 2032 students, 1209 students responded at every time-period in which they were included in the study. At the other extreme, only 85 failed to respond at any time-period, and these were omitted from further consideration. Based on the intended sample, the participation rates were, at each wave, respectively, 90, 85, 84, 80, 78 and 75 per cent, indicating some potential for response bias, especially in later waves, despite a generally high rate of follow-up for this type of study.

At each wave, a questionnaire was administered by lap-top computer.[8] Where a subject was unavailable at school, from wave 4 on, the questionnaire was administered by telephone. The questionnaire was presented as dealing with important health issues for adolescents and included questions on a wide range of health risk behaviours and mental health.

A principal focus of the analysis has been on patterns of cigarette smoking (since this is the single behaviour most likely to impair the long-term health of adolescents) and this tutorial focuses on one of the smoking outcome measures, 'daily' or 'regular' smoking. On each occasion a subject's self-assessed smoking status was determined using a 7-day retrospective diary, completed by all subjects except those who considered themselves non-smokers or ex-smokers (no cigarette in the previous month). From the diary response, a subject was categorized as a 'daily (regular) smoker' if they reported smoking on at least six days of the previous week.

Demographic and family variables were assessed at entry to the study, including date of birth, sex, country of birth, parental education and parents' marital status. For each subject an indicator of parental smoking was based on whether at least one parent was reported to be a daily smoker. At each wave, participants also answered a question about the extent to which their peers smoked. A measure of mental health status was obtained using a computerized form of the revised Clinical Interview Schedule (CIS), which is designed to assess symptoms of depression and anxiety in non-clinical populations.[9] This instrument generates a total score which was categorized into four levels of psychiatric morbidity.

The primary questions that we sought to answer with respect to patterns of smoking in the adolescent population were:

(i) What is the prevalence of smoking for each gender, how does it change with age and how does it relate to baseline factors such as parental smoking? (The latter part of this question will not be considered in this tutorial but the same methods apply.)
(ii) What is the incidence of new cases of smoking (and the incidence of quitting smoking among smokers) and how does this depend on age, sex and other factors including time-dependent measures such as mental health status?

There is a clear conceptual distinction between these two questions. The first relates to average patterns in the population and could in principle be studied with cross-sectional data, assuming that trends with calendar time are minimal. The second question is inherently longitudinal and cannot be addressed without considering patterns of change *within individuals.*

The statistical analysis aims to produce valid answers to these questions, taking proper account not only of the longitudinal nature of the data but also of the way in which the cohort was assembled (using a two-stage stratified survey design) and the fact that there are missing data.

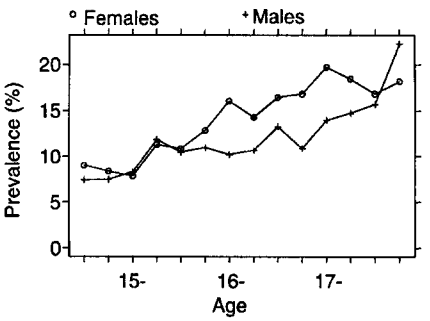


Figure 2. Prevalence of daily smoking by age (crude proportions)

3. PREVALENCE ANALYSIS

Our discussion concentrates on a binary outcome variable, Y_{ij}, which represents self-reported daily smoking (1 = yes, 0 = no), where i indexes subjects ($i = 1, \ldots, n$) and j indexes occasions ($j = 1, \ldots, J$). In this study there were six planned occasions of measurement so $J = 6$, although for most subjects at least one value was missing, either by design (in that they were not included for the first wave) or by 'happenstance' in that they failed to complete the survey on one or more occasions. In studying the prevalence of smoking, we are interested in an underlying population quantity that may be a function of a number of covariates such as age, sex and parental smoking.

For simplicity we consider a model relating smoking prevalence to age and sex alone. Figure 2 shows a graph of the crude prevalence of daily smoking by sex and by age in 3-month intervals, the age interval chosen to ensure that the same individual cannot contribute more than once to any one point in the graph. Note however that the dependence of smoking prevalence on age illustrated in this graph may reflect both cross-sectional age differences and longitudinal trends. Age was chosen for this initial analysis because of its intrinsic epidemiological importance, but we will also discuss an analysis using study wave in place of age.

The natural statistical model for investigating the dependence of smoking prevalence on age and sex is the logistic regression model:

$$\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 \text{sex} + \beta_2 \text{age} + \beta_3 \text{sex} \times \text{age} \tag{1}$$

where sex is coded as 0 for males and 1 for females, age is measured in years, and age × sex indicates the interaction. One of the aims of the study was to investigate the extent to which any age-related increase in smoking prevalence was different between the sexes, hence our inclusion in the model of β_3, which can be interpreted as the difference in age effect between females and males, while β_2 is the age effect for males.

Despite the fact that we have longitudinal data it is important to remember that this is simply a model for the population prevalence (a cross-sectional quantity) corresponding to a particular age and sex. This is sometimes called a marginal model because it does not specify a full probability model for the outcome variables, Y_{ij}, considered jointly. In order to fit the model (that is, obtain appropriate estimates of the parameters $\boldsymbol{\beta} = \{\beta_0, \beta_1, \beta_2, \beta_3\}$), we need either to specify

a full probability model for the available data y_{ij}, from which an appropriate likelihood function could be derived, or alternatively to decide on a reasonable estimation method (in the absence of a full model). If we attempt to specify a full probability model we are faced immediately with the problem of dependence or correlation between values of y_{ij} on the same individual. Before addressing this problem we review the method of maximum likelihood estimation assuming independent observations, since this provides a helpful reference point for the later methods.

If there were no correlation within individuals, we could reasonably assume that

$$y_{ij} \sim \text{Binomial}(1, \pi_{ij}), \text{ independently for all } i, j$$

and derive the likelihood function, up to a proportionality constant, as

$$\prod_i \prod_j \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{(1 - y_{ij})}.$$

To find maximum likelihood estimates of $\boldsymbol{\beta}$ it is simpler to work with the log-likelihood

$$l = \sum_i \sum_j [y_{ij} \log \pi_{ij} + (1 - y_{ij}) \log (1 - \pi_{ij})]. \tag{2}$$

Maximizing this requires finding the value of $\boldsymbol{\beta}$ for which the derivatives of (2) with respect to the components of $\boldsymbol{\beta}$ are zero and it is a simple standard exercise to show that this amounts to solving simultaneously the (four) equations

$$U(\beta_k) = \frac{\partial l}{\partial \beta_k} = \sum_i \sum_j x_{ijk} (y_{ij} - \pi_{ij}) = 0 \tag{3}$$

where k indexes the four β parameters and x_{ijk} is the corresponding covariate value (with $x_{ij0} = 1$) for individual i on occasion j. These equations are known as the likelihood score equations and for the logistic regression model (1) they can be written in the compact vector form

$$\mathbf{U}(\boldsymbol{\beta}) = X^{\mathrm{T}}(\mathbf{y} - \boldsymbol{\pi}) = 0 \tag{4}$$

where $\mathbf{y}$ and $\boldsymbol{\pi}$ are vector forms of the data y_{ij} and parameters π_{ij}, respectively, and X is the corresponding 'design matrix' with number of rows equal to the length of the $\mathbf{y}$ vector and number of columns equal to the number of $\boldsymbol{\beta}$ parameters. The solutions of the score equations are the maximum likelihood estimates, $\hat{\boldsymbol{\beta}}^{\mathrm{ML}}$, and a well known iterative generalization of the method of weighted least squares provides an efficient numerical algorithm for obtaining these estimates,[10] available in many modern statistical packages, such as GLIM, SAS, S-plus and Stata. Under the assumption of independent observations the asymptotic covariance matrix of these estimates is given by the information matrix

$$\text{cov}(\hat{\boldsymbol{\beta}}^{\mathrm{ML}}) = (X^{\mathrm{T}} \hat{A} X)^{-1} \tag{5}$$

where $\hat{A} = \text{diag}\{\hat{\pi}_{ij}(1 - \hat{\pi}_{ij})\}$, a diagonal matrix containing the binomial variances calculated at the values of π_{ij} corresponding to the ML estimates $\hat{\boldsymbol{\beta}}^{\mathrm{ML}}$.

3.1. Improved standard errors for ML estimates

Just how wrong is the maximum likelihood (ML) method likely to be with our longitudinal data, where repeated responses on the same individual are very likely to be correlated? First, it is easy to see from (4) that solving the likelihood-based score equations will give *consistent* estimates of $\boldsymbol{\beta}$ regardless of the full probability model (with its potential dependencies), as long as the

first-order specification of the mean or prevalence (1) is correct. In other words, with large samples, the point estimates should be close to the true population values. Second, however, it is likely that the estimated standard errors obtained from (5) will be too small (assuming the correlation within subjects is positive), since they are based on assuming we have more independent data points than we really do. A first approach to improving on the use of maximum likelihood is then to stick to the point estimates but to 'fix up' the standard errors. Several methods are available for doing this, each relying on various asymptotic approximations (meaning that they will be unreliable if sample size is small):

(i) *Bootstrap.* The idea underlying the bootstrap is to resample from the observed data, drawing with replacement to achieve a sample of the same size each time, and to use the variation in the estimated parameter across the set of bootstrap samples in order to estimate the sampling variability of the estimate.[11] With correlated data as in the present example, the bootstrap sampling should draw with replacement from the set of independent subjects, so that intra-subject correlation is preserved in the bootstrap samples.

(ii) *Jack-knife.* In our application, using the jack-knife would involve dropping each subject in turn from the total cohort and estimating model parameters from the remaining subjects' data. Repeating this for each subject results in n different sets of estimates. The variation between these estimates can be used to obtain an approximate covariance matrix for $\hat{\boldsymbol{\beta}}^{\mathrm{ML}}$.

(iii) *Information-sandwich (Huber/White) estimate of variance.* Both the bootstrap and jack-knife techniques are computationally intensive. On the other hand, the information-sandwich method[12-14] involves a closed-form calculation, based on an asymptotic (large-sample) approximation, and still provides good results in many situations. A heuristic explanation is provided in the Appendix. For the current model the resulting formula, providing a 'robust' variance estimate (that is, consistent under misspecification of the dependence structure), is

$$\mathrm{cov}_{\mathrm{R}}(\hat{\boldsymbol{\beta}}^{\mathrm{ML}}) = (X^{\mathrm{T}}\hat{A}X)^{-1}\sum_{i=1}^{n}(X_i^{\mathrm{T}}(\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i)^{\mathrm{T}}X_i)(X^{\mathrm{T}}\hat{A}X)^{-1} \tag{6}$$

where the summation is over the unique individuals and X_i, $\mathbf{y}_i$ and $\hat{\boldsymbol{\pi}}_i$ are the individual-specific (matrix/vector) components of X, $\mathbf{y}$ and $\hat{\boldsymbol{\pi}}$. A generic facility for calculating robust variance formulae such as these is available within the Stata package,[3] versions are also available in other software such as SAS and SUDAAN.[15]

3.2. Improving the estimates as well as the standard errors: quasi-likelihood and GEE

While the naive maximum likelihood estimates obtained by solving (4) are consistent given only the specification of the first-order model (1), they are not as *efficient* as estimates from a method that more fully utilizes information on the data's structure, including dependencies over time. A method based on maximizing the likelihood under a full probability model would be (asymptotically) the most efficient, but a useful method that does not require the specification of a full model yet still offers greater efficiency is the *quasi-likelihood* approach. Quasi-likelihood estimates are obtained in general by solving the equations

$$\mathbf{U}^{\mathrm{q}}(\boldsymbol{\beta}) = D^{\mathrm{T}}V^{-1}(\mathbf{y} - E(\mathbf{y})) = 0 \tag{7}$$

where the $N \times K$ matrix D (N being the overall length of the data vector $\mathbf{y}$ and K the length of the parameter vector $\boldsymbol{\beta}$) contains the derivatives $\partial E(y_{ij})/\partial \beta_k$, and $V = \text{cov}(\mathbf{y})$ is the covariance matrix of the y_{ij}'s, which contains known functions of $E(y_{ij})$ and possibly other unknown parameters. Note that, for our logistic regression model, $E(y_{ij}) = \pi_{ij} = (1 + e^{-\mathbf{X}_{ij}'\boldsymbol{\beta}})^{-1}$, where $\mathbf{X}_{ij}$ denotes the column vector of covariate values x_{ijk}. These estimating equations are the weighted least squares 'normal equations' that result from minimizing $(\mathbf{y} - E(\mathbf{y}))V^{-1}(\mathbf{y} - E(\mathbf{y}))$ when V contains known constants. Where V is functionally dependent on $E(y_{ij})$ (that is, on $\boldsymbol{\beta}$) and possibly on other parameters, the solution of these equations has been shown to have a number of attractive properties.[10,16] The appeal of the quasi-likelihood approach is that it requires only the specification of second-order structure, in the form of the variance V, but does not demand a full model.

For our logistic regression specification, the estimating equations (7) become

$$X^{\mathrm{T}}AV^{-1}(\mathbf{y} - \boldsymbol{\pi}) = 0. \tag{8}$$

Clearly if $V = A$ the equations reduce to the independence ML equations. More generally, the numerical solution of the 'score-type' estimating equations (7) is readily accomplished using the same iteratively reweighted least squares algorithm (IRLS) used for the likelihood equations.[10]

3.2.1. Generalized estimating equations

In the context of repeated measurements (clustered or longitudinal) on $i = 1, \ldots, n$ individuals, a generalization of quasi-likelihood that has become widely used is the *generalized estimating equations* (GEE) method.[5,17] The key to GEE is the representation of V in the estimating equations (7) as a block diagonal matrix consisting of n submatrices V_i having the form

$$V_i = A_i^{1/2} R(\boldsymbol{\alpha}) A_i^{1/2} \tag{9}$$

where A_i is the diagonal submatrix of A corresponding to individual i and $R(\boldsymbol{\alpha})$ is termed a 'working' correlation matrix and is a function of further unknown parameters $\boldsymbol{\alpha}$. The role of the working correlation matrix is to provide a guess at the true marginal covariance matrix $V(\mathbf{y})$ and it turns out that as long as the robust information-sandwich method is used for standard errors, the GEE method works well in large samples even if $R(\boldsymbol{\alpha})$ is misspecified. Incorporating a guess at the correlation structure increases the efficiency of estimation of $\boldsymbol{\beta}$ (that is, the variance of the estimate should be smaller). If we set $R(\boldsymbol{\alpha}) = I$, the identity matrix, then once again we recover the ML estimates. More generally, for an initial value of $\boldsymbol{\alpha}$, IRLS provides an initial estimate of $\boldsymbol{\beta}$ by solving (7). Given this $\boldsymbol{\beta}$ the 'GEE method' is to construct standardized residuals for the y_{ij}'s from which an updated estimate of $\boldsymbol{\alpha}$ is obtained (using the method of moments). These iterations are repeated until convergence.

If the first and second moment specifications are correct, the formal parallel between quasi-likelihood and likelihood can be used to show that large-sample standard errors for the GEE estimates, $\hat{\boldsymbol{\beta}}^{\text{GEE}}$, are provided by the 'quasi-information' matrix

$$\text{cov}_{\text{M}}(\hat{\boldsymbol{\beta}}^{\text{GEE}}) = (D^{\mathrm{T}}V^{-1}D)^{-1} \tag{10}$$

where the subscript 'M' stands for 'model-based', since the formula will not produce consistent estimates if the assumed form of V is wrong. In practice, of course, estimates of D and V based on substituting the estimated values of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ must be used. Again, however, the

information-sandwich idea can be used to 'robustify' these standard errors, leading to

$$\mathrm{cov}_{\mathrm{R}}(\hat{\boldsymbol{\beta}}^{\mathrm{GEE}}) = (D^{\mathrm{T}}V^{-1}D)^{-1}\sum_i (D_i^{\mathrm{T}}V_i^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i)^{\mathrm{T}}V_i^{-1}D_i)(D^{\mathrm{T}}V^{-1}D)^{-1} \tag{11}$$

which is simply a generalized version of the previous formula (6), which robustified the standard error of the independence ML estimate.

To capitalize on the possible efficiency gains of the GEE method, appropriate specification of $R(\alpha)$ is required. In some situations a single-parameter 'exchangeable' model, in which $\mathrm{corr}(Y_{ij}, Y_{ik}) = \alpha$ for all $j \neq k$ may provide substantial improvements over the independence assumption; for longitudinal data an autoregressive or other 'banded' correlation structure may be more appropriate. With large n and small J an unstructured correlation with $J(J-1)/2$ parameters may be feasible and provides maximum flexibility.

3.3. A missing-data adjustment

None of the previously described methods produces valid estimates if there are missing data, unless the data are missing 'completely at random',[18] meaning that the chance that a particular response value is missing is independent of the observed responses and covariates for that subject. Heuristically, the methods break down because the estimating equations on which they are based no longer have zero expectation if there are subjects missing in a non-random fashion, so that consistency of the resulting estimates cannot be guaranteed. This view of the problem suggests a simple method of modifying the estimating equations by applying weights to allow for the pattern of missingness.[19-22] We use a weighted version of (3):

$$U^{\mathrm{w}}(\beta_k) = \sum_i \sum_j w_{ij}U_{ij}(\beta_k) = \sum_i \sum_j w_{ij}x_{ijk}(y_{ij} - \pi_{ij}) = 0 \tag{12}$$

where w_{ij} is a weight equal to the inverse probability that the jth response for individual i is observed (supposing for the moment that these probabilities are known). These weights can be expressed more formally as $w_{ij} = 1/\Pr(y_{ij}\ \text{observed}) = 1/E(I_{ij})$, where I_{ij} denotes a random variable that takes the value 1 if y_{ij} is observed and 0 if not. The weighted estimating equations are unbiased under repeated sampling of the y's, assuming the same sampling probabilities or weights, since

$$\begin{aligned} E_I[E_y(U^{\mathrm{w}}(\beta_k))] &= E_I\left[E_y\left(\sum_i \sum_j I_{ij}w_{ij}U_{ij}(\beta_k)\right)\right] \\ &= E_y\left[\sum_i \sum_j E_I(I_{ij})w_{ij}U_{ij}(\beta_k)\right] \\ &= E_y\left[\sum_i \sum_j U_{ij}(\beta_k)\right] = 0 \end{aligned}$$

(where $E_Z(.)$ denotes taking the expectation under the distribution of the random variable Z). The w_{ij} serve to weight the score equations in such a way that we compensate for missing data values.

Weighting the estimating equations in this way is formally similar to quasi-likelihood in that the weighted estimating equations can be written in exactly the same form as (8) with AV^{-1} replaced by W, a diagonal matrix containing the weights w_{ij}. The parallel with quasi-likelihood again indicates that the same numerical methods can be used to solve the weighted estimating

equations, and similarly a robust variance estimate allowing for possible dependence within individuals can be obtained. It should be noted, however, that the weighted estimating equations do not incorporate a within-subject correlation structure in the way that the GEEs do. It would be attractive to combine these two features into the estimating equations, but it is not clear how to handle the estimation of the α parameters when weights are present.

In practice, the probability of being observed is not known, so we need to estimate the weights in some reasonable way.[23] We do this using a method related to that of 'post-stratification' in the survey sampling literature.[24] An initial analysis was performed using logistic regression to determine which of a number of fixed covariates were most strongly predictive of subjects completing only one or two waves. The covariates identified were sex, country of birth (Australia or not), parental divorce and smoking status at recruitment. These four dichotomous variables created a cross-classification of subjects into 16 cells. At each wave empirical weights were estimated within each of these cells as the inverse of the response rate achieved within that cell. Note that these weights will not be reliably estimated unless reasonable numbers are available within each cell. Estimation of the weights implies that the consistency argument given above no longer strictly holds, but it still provides an appealing heuristic justification for the method, assuming that the post-stratification is in fact predictive of the missing data mechanism. The method is valid as long as the missing data are a random subsample of responses within the cells of the cross-classification.

3.4. Allowing for sample survey design

A final consideration that may be accommodated within the marginal modelling analysis is the complex design of the survey from which the data arose. Recall that the initial cohort was identified from a two-stage sampling process where schools were selected by stratified random sampling at the first stage. Full consideration of the issues raised by sample survey design is beyond the scope of this tutorial, partly because these issues rapidly become tangled with philosophical questions about the nature of statistical inference concerning finite populations, the traditional starting point of sample survey theory.[25] We retain a pragmatic view for the current exposition and simply point out that the recruitment of a study group via a known sampling process in a finite population may involve one or all of the following:

(i) sampling weights, which reflect differential probabilities of being included for different members of the population;
(ii) stratification, where sampling was carried out independently within sections (strata) of the population;
(iii) clustering, where a multi-stage selection process may produce a sample with higher levels of correlation within identified subgroups or clusters of the sample than between such subgroups.

The effect of these features can be accommodated within the estimating equation framework. In particular, the sampling weights can be handled in exactly the same way as described in the previous section. Once the weights have been allowed for, stratification and clustering do not affect the expected value of the 'pseudo-score' statistic and so the estimating equations continue to provide consistent estimates. The sampling design may, however, affect the standard errors, in that stratification will lead to more precise estimation if between-strata variance is larger than

within, while clustering may lead to less precise estimation because of dependencies within clusters. These effects may be allowed for in the information-sandwich approximation.[3,19,25] First, the effect of stratification is dealt with in the calculation of the empirical variance of the observed y_{ij} that is used in the 'middle' of the sandwich, using standard sample survey concepts of variance estimation under stratification. Second, the clustering may be allowed for by expanding the unit on which the empirical variance of the sandwich formula is based from the individual level to the cluster (school) level. Thus, ignoring the other issues, allowance for cluster effects could be achieved in (6) by replacing the summation over i with a summation over (say) s, where s indexes schools, and the corresponding X_s, $\mathbf{y}_s$ and $\hat{\boldsymbol{\pi}}_s$ are the school-specific components of X, $\mathbf{y}$ and $\hat{\boldsymbol{\pi}}$.

3.5. Results

Model (1) was fitted to the adolescent cohort data using the succession of methods discussed above. The analysis was carried out using the statistical package Stata[3] and we present Stata commands as they were used with our data. Note that lines beginning with '*' are comments in Stata. The following labelling commands serve to introduce the variables used in the analysis:

```
* Label variables for prevalence analysis
* (NB Stata variable names are limited to 8 characters)
label variable id          "identification of individuals"
label variable wave        "Wave of data collection: 1 to 6"
label variable regsmoke    "Indicator of regular smoking"
label variable c_age       "Age centred at mean of cohort"
label variable wgt_mv      "Missing data weights"
label variable wgt_str     "Strata sampling weights"
label variable sregion     "Stratum to which school belongs"
```

The first method of fitting model (1) was maximum likelihood, which corresponds to solving the estimating equations (4), with model-based standard errors given by the inverse information matrix, (5). This is accomplished in Stata by the logit (or closely related logistic) command, which is a special case of the glm command for performing ML estimation in generalized linear models. The command is coupled with the xi command, which generates dummy and interaction variables:

```
* Method (1): Maximum Likelihood logistic regression with model-based SEs
xi: logit regsmoke i.sex*c_age
```

The results from this and subsequent methods are presented in Table I, where we present age effects for males and females separately, that is, $\hat{\beta}_2$ and $\hat{\beta}_2 + \hat{\beta}_3$, respectively. The latter effect is not standard output from Stata regression commands but can be obtained after fitting the model, by using the lincom command, which obtains the standard error of linear combinations of estimated parameters, using the estimated variance-covariance matrix of the vector $\boldsymbol{\beta}$:

```
* Obtain age effect in females = age effect in males + interaction effect
lincom _b[c_age] + _b[IsXc_a_1]
* Stata's xi command automatically generates the name IsXc_a_1 for the
* interaction variable sex.age
```

The second method of fitting the model, maximum likelihood with robust standard errors, also solves the estimating equations (4) to estimate the model parameters, but standard errors are

Table I. Parameter estimates (with standard errors) from six different methods of estimating the expanded version of model (1) that includes the interaction between sex and age*

	Maximum likelihood (ML)				Survey estimation weighted for strata and missingness (Section 3.4)	
	1 Model-based SEs (Section 3)	2 Robust SEs (Section 3.1)	3 GEE (unstructured correlation) (Section 3.2)	4 ML robust weighted for missingness (Section 3.3)	5 Individual as psu	6 School as psu
Constant†	−2·08		−1·94	−2·00	−1·94	
	(0·050)	(0·085)	(0·082)	(0·086)	(0·095)	(0.110)
Sex (female)	0·25		0·18	0·23	0·18	
	(0·066)	(0·117)	(0·113)	(0·118)	(0·129)	(0·140)
Age (per year):						
males	0·27		0·28	0·30	0·26	
	(0·059)	(0·060)	(0·046)	(0·060)	(0·062)	(0·058)
females	0·33		0·33	0·35	0·33	
	(0·050)	(0·052)	(0·037)	(0·052)	(0·061)	(0·041)

* The analysis was restricted to 1867 adolescents who were no more than 1 year older or younger than the mean age of the cohort at the second wave, in order to reduce potential influence of individuals of relatively unusual age

† Age was centered at the mean of the cohort, 16·2 years, so the constant estimates the log-odds of smoking in males (the reference category of the sex effect) at this age.

obtained by the information sandwich formula (6):

```
* Method (2): Maximum Likelihood logistic regression with robust SEs
xi: logit regsmoke i.sex*c_age, cluster(id)
lincom _b[c_age] + _b[IsXc_a_1]
```

The model was next fitted using GEE, that is, solving the estimating equations given by (8) with an unstructured working correlation matrix, $R(\alpha)$, defined by corr $(Y_{ij}, Y_{ik}) = \alpha_{jk}$ for all $j \neq k$ and using (11) to estimate robust standard errors:

```
* Method (3): GEE with unstructured working correlation matrix
xi: xtgee regsmoke i.sex*c_age, family(binomial) correlation (unstructured) i(id) t(wave) robust
lincom _b[c_age] + _b[IsXc_a_1]
```

To fit the model with a weighting mechanism for missing data, solving equations (8) again provided the parameter estimates, and formula (11) the standard errors, but in this case AV^{-1} was replaced by W as discussed in Section 3.3:

```
* Method (4): ML with robust SEs. Weighted for missing data
xi: logit regsmoke i.sex*c_age [pweight = wgt_mv], cluster(id)
lincom _b[c_age] + _b[IsXc_a_1]
```

The weights used in this calculation (variable 'wgt_mv') were calculated in a spreadsheet and merged into the main data set.

The fifth and sixth methods of fitting model (1) weighted the estimating equations (8) to take account of the sampling scheme as well as the missing data (Section 3.4). Stata provides specialized 'survey estimation' commands for this purpose (where 'svylc' is equivalent to 'lincom' used above):

```
* Method (5): Survey estimation with individual as primary sampling unit
*
* First calculate the weights for missing data and strata
generate wgt_both = wgt_mv*wgt_str
* Then fit the model
xi: svylogit regsmoke i.sex*c_age [pweight = wgt_both], psu(id) strata (sregion)
svylc _b[c_age] + _b[IsXc_a_1]
*
* Method (6): Survey estimation with school as primary sampling unit
*
xi: svylogit regsmoke i.sex*c_age [pweight = wgt_both], psu(school)
svylc _b[c_age] + _b[IsXc_a_1]
```

Several interesting features emerge from a comparison of the results obtained under the different methods (Table I). First, the effect of 'robustifying' the standard errors for the ML estimates is minimal except for the 'sex' effect (note that age was centered in these calculations so the 'sex' effect represents the estimated difference in log-odds at the mean age). The increase in standard error reflects the fact that 'sex' is a between-subject effect, in contrast to 'age', whose variation in these data is primarily within-subject. The variance of between-subject effects is especially likely to be underestimated by the naive ML approach, since by ignoring within-subject correlation it

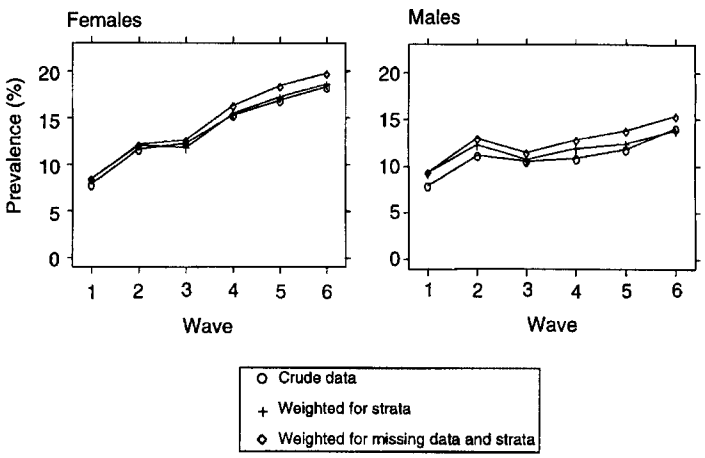


Figure 3. Prevalence of daily smoking by wave

effectively exaggerates the amount of between-subject information in the data. Proceeding to the GEE estimates, we see a small increase in the estimated intercept, which corresponds to the log-odds of smoking among males at the mean age, and a fall in the estimated sex effect. Standard errors of the between-subject effects (intercept and sex effect) are again larger than the naive estimates, but for the within-subject age effect the standard errors are reduced, reflecting the increased efficiency gained by recognizing intra-subject correlation in the estimating equations.

The weighted estimating equation methods (4–6 in Table I) give generally similar coefficient estimates to the GEE method, although method 4, which only weights for missingness, is less similar than methods 5 and 6 where stratum weights are also included. The patterns of missing data were such that the results of methods 4–6 can be seen empirically to compensate for missingness, since there was a tendency for more males and more smokers to be missing, especially at the later waves. Compared with the males, there is a smaller difference in females between the weighted and unweighted estimates of smoking prevalence at the mean age. These facts are illustrated in Figure 3, which displays prevalence estimates by wave (corresponding approximately to 6 month age increments), and compares crude frequencies with those obtained after (i) weighting for strata and (ii) also weighting for missingness. These prevalence estimates were obtained using Stata's command for survey estimation of means, which incorporates weights using the same principles as already used in the logistic regression:

```
* Figure 3
* Clear existing survey estimation settings
svyset, clear
*
* Raw proportions of regular smoking in each wave for both sexes
svymean regsmoke, by(sex wave)
*
* Proportions weighted for strata
svymean regsmoke [pweight = wgt_str], by(sex wave) strata(sregion)
*
* Proportions weighted for strata and missing data
svymean regsmoke [pweight = wgt_both], by(sex wave) strata(sregion)
```

Finally it is worth noting (comparison of methods 5 and 6) that the standard errors using the robust information sandwich based on empirical variance estimates within *schools* (that is, allowing for possible intra-school correlation effects) are little different from those based on individuals, suggesting that intra-school correlation is negligible for these parameter estimates.

There is little variation across methods in the estimated age effects for either males or females. Both suggest a highly significant increase in prevalence with age, at a rate of about 0·3 (log-odds) per year, corresponding to an increase in the odds of daily smoking of about $e^{0·3} = 1·35$ fold per year (equivalent to an increase in the *risk* per year of about 1·3 for baseline risk in the range 10–15 per cent). There is a slight suggestion that the rate of increase of males may be lower than females.

Different conclusions emerge from Table II which shows the results of a similar analysis of daily smoking prevalence in which instead of using each student's actual age in the model we use the wave of the study (rescaled to units of years so that coefficient estimates are comparable with those in Table I):

```
* Table 2: Analysis by wave
* All commands follow age analysis with c_age replaced by c_wave
generate c_wave = (wave - 3·5)/2
* c_wave is centered between waves 3 and 4 and has increments of 0·5
```

Using wave as a surrogate for age ensures that rates of change with time now reflect longitudinal changes within the cohort and are not confounded with possible cross-sectional age differences. In fact in this model, the rate of increase with wave among males is about half that among females, with the difference between the two statistically significant under model 5 at $p = 0·003$ (using the Wald-type test based on comparing the ratio of the estimated interaction effect to its standard error with the standard normal distribution). Why does the analysis by wave indicate a much lower rate of increase in daily smoking with age among males? The explanation is clear from Figure 4, which plots age trends (superimposed lines) in the cohorts of males and females at each wave of the study:

```
* Figure 4
label variable centile "Centiles of age at each wave"
* (Tertiles were defined for wave 1 and sextiles for all subsequent waves)
svymean regsmoke [pweight = wgt_both], by(sex wave centile) strata(sregion)
```

It can be seen that there is a consistent age-related increase *within* the cohort of males at each timepoint; for the females such an increase appears in a weaker form only in the middle of the study. This suggests an interesting substantive conclusion, that the males are more strongly differentiated on an age basis and remain so throughout the study, whereas the females' smoking behaviour appears to be more homogeneous with respect to inter-individual age differences but increases more sharply over time.

Apart from the sex difference in age/wave effect and the general patterns of difference across methods already described in relation to Table I, the main distinctive feature of the results of Table II is the large discrepancy between the GEE-based estimate of the male wave effect (column 3) and the strata-and-missingness weighted estimate in columns 5 and 6. Theoretically, the GEE estimate may have an efficiency advantage (and indeed its standard error is smaller) but this appears to be considerably outweighed in these data by bias effects due to sample selection and missing data, which the GEE method does not allow for. Finally it is interesting to note that the

Table II. Parameter estimates (with 95 per cent confidence intervals) for the model of Table I with wave of study used as a predictor in place of age*

	Maximum likelihood (ML)				Survey estimation weighted for strata and missingness (Section 3.4)	
	1 Model-based SEs (Section 3)	2 Robust SEs (Section 3.1)	3 GEE (unstructured correlation) (Section 3.2)	4 ML robust weighted for missingness (Section 3.3)	5 Individual as psu	6 School as psu
Constant†	− 2·08		− 1·95	− 2·00	− 1·93	
	(0·050)	(0·085)	(0·082)	(0·086)	(0·097)	(0·113)
Sex (female)	0·23		0·17	0·20	0·14	
	(0·225)	(0·117)	(0·114)	(0·119)	(0·130)	(0·143)
Wave (per year):						
males	0·17		0·25	0·21	0·16	
	(0·062)	(0·057)	(0·047)	(0·057)	(0·057)	(0·061)
females	0·33		0·33	0·35	0·35	
	(0·053)	(0·043)	(0·037)	(0·042)	(0·046)	(0·038)

* Study wave was rescaled so that the corresponding coefficients represent approximate rate of annual increase. See Table I for details of cohort restriction

† Wave was centred between waves 3 and 4, so the constant estimates the log-odds of smoking in males (the reference category of the sex effect) at this point

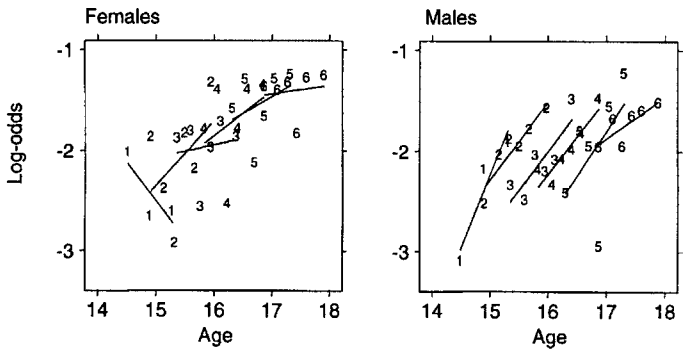


Figure 4. Log-odds of daily smoking by centile of age for each wave of study (estimates weighted for strata and missing data). Wave is indicated by number (1–6) and points are plotted at the mean age of each centile, with a least-squares regression line superimposed for each wave

similarity in trend over time in the crude prevalence (corresponding to the naive ML estimates) and the strata-and-missingness-weighted estimates for the males can be clearly seen in Figure 3, which shows that strata weighting has a substantial effect only in the early waves and missingness weighting is only influential in the later waves.

4. INCIDENCE ANALYSIS

To address questions relating to which individual factors might predict the uptake of smoking in individuals, the prevalence or marginal analysis is of little use, for two reasons. First, the existence of associations between time-varying variables, such as smoking and depression, at a cross-sectional level does not allow us to infer causality, since we cannot tell which of the attributes was present in the individual first or whether indeed both attributes appear simultaneously as a result of some third factor or set of factors. Second, and somewhat more subtly, estimates of association made on the basis of a marginal model cannot be interpreted at a subject-specific level. For instance, in the models fitted in the previous section (Table II), the rate of increase in the log-odds of daily smoking for every year of increased age (that is, two waves of the study) among females was estimated to be 0·35. This means that, other things being equal, the log-odds of smoking in a population of females aged 16 is about 0·35 greater than that in a population aged 15, but it does not mean that for a randomly chosen individual female, the odds of being a smoker on her sixteenth birthday is $e^{0·35}$ greater than it was on her fifteenth. This is the distinction between subject-specific and marginal associations that many authors have described,[26,27] a distinction that does not arise in *linear* regression models.

In order to study questions relating to the development of behaviours over time we must go beyond marginal model specifications such as (1). We have a choice of going either to a global model, which specifies a pattern of change over time, or a local model, which specifies only the relationship between successive timepoints or occasions of measurement. The former type of model will generally involve a hierarchical structure based on a model for variation between subjects, where the mean value may exhibit some interpretable form (such as linear increase with time), and a model for variation within subjects. Such models go under a wide variety of names including (latent) growth curve,[28,29] multi-level,[30] hierarchical[31] and mixed (fixed and random

effects) models,[32,33] but only recently has substantial work been done on applying these sorts of models to binary outcomes.[34,35] In this section we describe a local or 'transitional' model, which aims to model the *incidence* of events that represent a change in the individual's status on an outcome measure between one occasion and a subsequent one.

Our analysis is limited to outcome events that only occur once for each subject: truly incident events in the usual epidemiological definition. In considering the incidence of regular smoking we therefore include only those subjects who had no history of regular smoking at the inception of the cohort. We may then consider our data in the form of an inception time, t_i^0, at which the subject was not smoking and a time of last review, t_i^f at which the subject may or may not have been smoking. The time scale used is the subject's age, the only scale that has a common meaning across all subjects, but of course each subject's status is only recorded at times that correspond to waves of the survey. A simple empirical analysis may be performed by calculating incidence rates from these data within subgroups defined by sex and other covariates, by the standard epidemiological method of dividing total number of events by total person-years of follow-up accumulated over all subjects in the subgroup. In this calculation, where an individual takes up smoking during their interval of observation, we assume that the person-years of follow-up is $t_i^f - t_i^0$ minus half the length of the last between-wave interval, since we do not have precise information on the time of incidence.

In order to examine more systematically the effect of risk factors and covariates on incidence, we seek an appropriate model to describe the rate of occurrence of the uptake of smoking between t_i^0 and t_i^f, allowing for differences in this time interval between subjects (most obviously, the interval will tend to be shorter for those who do in fact commence smoking). The particular model we propose is a discrete version of the proportional hazards regression model that is commonly used in survival analysis,[36,37] where the outcome measure is the time to a particular event. The exact time of occurrence of events such as 'started smoking' is not recorded in studies such as ours since we only know the time interval during which the event occurred or failed to occur; such data are termed 'interval censored'.[37] Define p_i as the probability that subject i becomes a daily smoker in their interval of observation. Expressed in terms of the underlying 'survival' variable T_i (the time to 'failure' for individual i), we may write $p_i = \Pr\{T_i \in (t_i^0, t_i^f) | T_i \geqslant t_i^0\}$.

A useful representation of the probability distribution of a survival variable is the instantaneous rate of failure or hazard rate (probability of failing at time t given survival until that time). If we assume that the underlying incidence process fits a proportional hazards model then the hazard rate for subject i, which we denote $\lambda_i(t)$, depends in a log-linear fashion on subject factors X_i, *independently* of time t:

$$\log \lambda_i(t) = \log \lambda_0(t) + X_i^T \boldsymbol{\beta} \tag{13}$$

where $\lambda_0(t)$ is a baseline hazard rate (applying to those individuals for whom $X_i = 0$). It can then be shown straightforwardly[36] that a particular transformation of p_i, the complementary log-log, also follows a linear model in X_i. Specifically

$$\log(-\log(1 - p_i)) = X_i^T \boldsymbol{\beta} + \log\left(\int_{t_i^0}^{t_i^f} \lambda_0(u)\, du\right). \tag{14}$$

The integral in this expression reflects the dependence of the baseline risk on the time interval (t_i^0, t_i^f) and, as long as $\lambda_0(t)$ does not vary greatly over the time span of interest, we may use the approximation $\int_{t_i^0}^{t_i^f} \lambda_0(u) du \approx (t_i^f - t_i^0)\bar{\lambda}_0$ where $\bar{\lambda}_0$ is the mean baseline hazard. This suggests that

much of the baseline effect may be captured by fitting the model

$$\log(-\log(1-p_i)) = \beta_0 + X_i^{\mathrm{T}}\boldsymbol{\beta} + \log(t_i^{\mathrm{f}} - t_i^0) \tag{15}$$

where $\beta_0 = \log(\bar{\lambda}_0)$. Our analysis of incidence proceeds by fitting the generalized linear model described by (15), using the method of maximum likelihood for binary outcomes,[10] which can be readily accomplished in a number of software packages such as SAS, GLIM and Stata.

The analysis just described assumes that the covariates represented by the vector X_i have a constant value over time for each subject, but a simple generalization allows for time-dependent covariates. We expand the data set to allow a separate record for each wave of data in each subject. Following the same logic as above, for every interval j (before the uptake of daily smoking), the corresponding probability p_{ij} can now be modelled as

$$\log(-\log(1-p_{ij})) = \beta_{0j} + X_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \log(t_{ij}^{\mathrm{f}} - t_{ij}^0) \tag{16}$$

where $t_{ij}^{\mathrm{f}} - t_{ij}^0$ is the length of the jth inter-wave interval for subject i and X_{ij} is the corresponding wave-specific covariate vector. The β_{0j} terms allow for possible variation in the baseline hazard across waves. Ignoring potential dependency within subjects a 'pseudo-likelihood' can be defined from the first-order specification (16) by the Bernoulli assumption that failure occurs with probability p_{ij}. As in the discussion of Section 3, solving the corresponding estimating equations will provide asymptotically unbiased estimates of the $\boldsymbol{\beta}$ coefficients, but the standard errors will be incorrect. Again, however, we can obtain asymptotically robust standard errors by using the information sandwich method.

4.1. Results

We initially calculated the observed incidence rates of regular smoking by dividing total number of events (a) by total person-years of follow-up (T) accumulated over all subjects in the subgroup. The rates in the different categories of the variables

```
label variable prev_smk  "Previous smoking status"
label variable CISscore  "Psychiatric morbidity"
label variable peer_smk  "Proportion of peers smoking"
label variable par_smk   "Parental smoking"
```

were calculated using Stata's 'survival time' commands

```
label variable start     "Indicator of initiation"
label variable Age_in    "Age at start of risk interval"
label variable Age_out   "Age at end of risk interval"
* Set the survival time (st) details for subsequent st commands
stset Age_out start, t0(Age_in) id(id)
strate prev_smk, scale(1000)
strate CISscore, scale(1000)
strate peer_smk, scale(1000)
strate par_smk, scale(1000)
```

and are presented per 1000 person years in Table III. The 95 per cent confidence intervals for these rates were calculated by the **strate** command[38] using standard approximations,[39] as $(a/T)\,e^{\pm 1.96\sqrt{1/a}}$. The covariate values were those reported by the respondent at the beginning of

Table III. Estimates* (with 95 per cent confidence intervals) of: 1, crude incidence rates per 1000 person years; 2, crude rate ratios (RR); 3, unadjusted RR from the complementary log-log survival model (16); 4, adjusted RR from model (18); 5, adjusted RR from model (16) with robust SEs

			RR from comp. log–log survival model		
	Incidence rate 1	RR (crude) 2	Unadjusted 3	Adjusted† 4 (model SE)	5 (robust SE)
Previous smoking status					
Non-smoker	17 (12, 24)	1 –	1 –	–	1 –
Ex-smoker	83 (58, 119)	4·9 (3·0, 8·0)	4·9 (3·0, 8·1)	(2·7, 7·5)	4·5 (2·7, 7·6)
None in last week	292 (238, 358)	17·3 (11·6, 25·7)	17·9 (12·0, 26·7)	(9·6, 22·4)	14·7 (9·3, 23·0)
1–4 days in last week	664 (517, 854)	39·4 (25·8, 60·1)	42·8 (27·9, 65·9)	(18·0, 45·9)	28·8 (17·5, 47·3)
Total CIS score					
0–5	59 (48, 71)	1 –	1 –	–	1 –
6–11	92 (69, 123)	1·58 (1·11, 2·23)	1·51 (1·07, 2·13)	(0·88, 1·81)	1·26 (0·87, 1·84)
12–17	107 (74, 154)	1·82 (1·20, 2·77)	1·80 (1·19, 2·73)	(0·82, 1·99)	1·28 (0·82, 2·01)
18 +	162 (120, 219)	2·77 (1·93, 3·97)	2·71 (1·89, 3·88)	(1·11, 2·45)	1·65 (1·05, 2·59)
Proportion of peers smoking					
None	31 (21, 46)	1 –	1 –	–	1 –
Some	60 (49, 74)	1·97 1·26, 3·08)	1·97 (1·26, 3·10)	(0·57, 1·45)	0·91 (0·56, 1·48)
Most	238 (196, 290)	7·75 (4·96, 12·10)	7·95 (5·08, 12·44)	(1·15, 3·06)	1·88 (1·12, 3·15)
Parental smoking					
Neither daily	63 (53, 75)	1 –	1 –	–	1 –
At least one parent daily	121 (99, 148)	1·92 (1·46, 2·51)	1·93 (1·47, 2·53)	(1·20, 2·09)	1·58 (1·19, 2·12)

* The analysis was restricted to 2-year age range, see Table I
† Adjusting for factors shown in addition to age and sex

the at-risk interval. The first three covariates – previous smoking status, psychiatric morbidity (measured by the CIS score) and peer smoking – were time-varying, while parental smoking was fixed at the response given in the first wave of data.

The ratio of the rates between two groups is estimated as $RR = \frac{a_1}{T_1}/\frac{a_2}{T_2}$ with corresponding 95 per cent confidence interval given by $RR \times e^{\pm 1.96\sqrt{(1/a_1 + 1/a_2)}}$. These rate ratios were obtained with the stmh command.[38]

```
* First obtain ratio of rate in category 2 of previous smoking status
* to rate in category 1 (previous "st" settings still in force)
stmh prev_smk, compare(2, 1)
* Similarly for ratio of rates in category 3 to 1, and 4 to 1
stmh prev_smk, compare(3, 1)
stmh prev_smk, compare(4, 1)
* Repeat for other covariates: CISscore, peer_smk, par_smk
```

When considered alone, each of the four factors analysed showed strong associations with the uptake of regular smoking, although a history of previous (occasional) smoking carried by far the strongest association.

We next employed model (16) with X_{ij}^{T} containing each covariate in turn:

```
generate risktime = Age_out-Age_in
label variable risktime "Time at risk in given wave"
* Fit the discrete-time proportional-hazards model with offset
xi: glm start i.prev_smk, family(binary) link(cloglog) lnoffset(risktime) eform
* Repeat replacing prev_smk with CISscore, peer_smk and par_smk in turn
```

This analysis gave us the unadjusted rate ratios from the survival model and these are shown in the third column of Table III. It is clear that the results (both point estimates and standard errors) are very similar to the empirically calculated rate ratios in column 2. To obtain the adjusted rate ratios we include all four covariates as well as sex and age, in X_{ij}^{T}.

```
* 1) Adjusted RRs from survival model (Model-based standard errors):
xi: glm start i.sex age i.prev_smk i.CISscore i.peer_smk i.par_smk,
                family(binomial) link(cloglog) lnoffset(risktime) eform
```

Having fitted this model in Stata, the robust standard errors can be calculated simply:

```
* 2) Adjusted RRs from survival model (Robust standard errors):
* using code available at http://ideas.uqam.ca/ideas/data/bocbocode.html
rglm, cluster(id)
```

The adjusted rate ratios with both model-based and robust standard errors are shown in the fourth and fifth columns of Table III. It can be seen that the effect of previous smoking history remains very strong while the strength of association of the other three factors is considerably weakened, especially that of peer smoking. These changes reflect predictable confounding effects; for instance, it is likely that adolescents with high levels of peer smoking will tend to have had more previous smoking experience that those with no peer smokers.

Finally, comparison of the last two columns shows the effect of using the robust information-sandwich standard error formula instead of the model-based method that erroneously assumes that all data points are independent. Only minor inflation of the confidence intervals was found.

5. DISCUSSION

In conclusion, we reiterate the importance of clearly defining the research questions of interest before undertaking detailed analysis of data collected from a panel or multiwave cohort study. In this tutorial we have limited our attention to the common problem of analysing a binary outcome

measure. When the outcome variable is continuous and modelled using normal distribution methods, the distinctions emphasized in this tutorial become somewhat less important. For answering questions concerning the *prevalence* of a disease or behaviour, one is concerned only with characterizing the first-order dependence of the outcome on covariates, and it is often possible to avoid complex modelling exercises which not only increase computational complexity but also raise the danger of results becoming sensitive to second-order assumptions in the modelling.

We have shown in a particular example that it is straightforward to estimate logistic regression parameters using estimating equations that include weights in order to adjust for known biases in the sample (due to patterns of missing data and survey design), and that the information-sandwich formula provides an attractive general way of estimating standard errors that are robust to misspecification of higher-order properties in large samples. The combination of weighted estimating equations and robust standard errors may be more attractive in many applications than the closely related GEE approach if the potential bias in (unweighted) parameter estimates is large relative to the precision that might be gained by allowing for the second-order structure in the estimating equations. Further study of the relative strengths of the alternative methods, perhaps by simulations in moderately sized samples, would be valuable. Our aim has been to illustrate the methods in practice and to explain both the practical details and the heuristic rationale. It should be emphasized that the estimating equation approaches and in particular the information-sandwich variance method are only guaranteed to work in large samples.

Most longitudinal studies will ultimately wish to address questions at a subject-specific level, in contrast to the population-averaged approach that suffices to examine prevalence-related questions. A similar distinction has been emphasized by others[6,27] but our specific approach to subject-specific modelling, based on modelling transitions or incident events, does not appear to have been widely used. We have adopted a standard proportional hazards survival model, which reduces to generalized linear modelling with a complementary log-log link function in the case of interval-censored data, to address questions about the relative importance of risk factors in the *incidence* of regular smoking. The technical details of this approach are not new, but it is difficult to find worked examples of the method in use.

The analyses presented do not by any means address all the questions that might legitimately be posed in relation to these data. First, it may be regarded as somewhat artificial to focus on transitions between 'non-daily smoking' and 'daily smoking'. A more reasonably conceptualization might be to think of subjects as having a more or less continuous underlying level of behaviour (in this case, smoking), whose level is imperfectly observed at the discrete occasions of the survey. One would envisage covariates acting on this underlying latent variable. Such a model would allow for measurement error in the outcome and would potentially be better able to distinguish true risk factors.

Finally, we have only partially dealt with the potential problems caused by missing data, by use of the method of empirical probability weighting in the prevalence analysis. This weighting method can only adjust for missingness that is predictable from observed covariates, that is, for data that are missing at random[18] (in our formulation, random within the cells for calculation of the empirical probability weights). Since the scope for missing data bias is generally less in an incidence than in a prevalence analysis, we have made no allowance for it in our incidence analysis. This could be examined in more detail, either building full probability models that include models for the missing data mechanism[18] and then using simulation-based methods for inference,[40] or possibly exploiting recent developments in semi-parametric modelling

approaches.[41] Neither of these possibilities is available in an accessible form with current computer software.

APPENDIX: THE INFORMATION-SANDWICH METHOD FOR ROBUST STANDARD ERRORS

The principle of the information-sandwich method is the same in each of its specific applications and we describe it in the context of the generic set of estimating equations given by the quasi-likelihood specification (7):

$$\mathbf{U}(\boldsymbol{\beta}) = D^{\mathrm{T}}V^{-1}(\mathbf{y} - \boldsymbol{\pi})$$

where $\boldsymbol{\pi} = E(\mathbf{y})$. As discussed, these reduce in special cases to the likelihood score equations (3) as well as the other specifications considered in Section 3. The solution $\hat{\boldsymbol{\beta}}$ of these equations provides unbiased estimates of the regression parameters, $\boldsymbol{\beta}$.

To derive an expression for the variance of the resulting estimates, the key preliminary step is to show that

$$-E\left(\frac{\partial \mathbf{U}}{\partial \boldsymbol{\beta}}\right) = D^{\mathrm{T}}V^{-1}D \tag{17}$$

a result that follows straightforwardly by interchanging the expectation (with respect to $\mathbf{y}$) and differentiation (with respect to $\boldsymbol{\beta}$) operators. The quantity (17) is known as the (expected) *information.*

Now a standard Taylor-series expansion of $\mathbf{U}(\hat{\boldsymbol{\beta}})$, considered as a function of the estimate $\hat{\boldsymbol{\beta}}$, about the true value $\boldsymbol{\beta}$, gives

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx -\left(\frac{\partial \mathbf{U}}{\partial \boldsymbol{\beta}}\right)^{-1} \mathbf{U}(\boldsymbol{\beta}).$$

This relationship forms the basis of successive updates of the estimate $\hat{\boldsymbol{\beta}}$ in the iteratively reweighted least squares estimation method, in which the gradient of $\mathbf{U}$ is replaced by its expected value, $D^{\mathrm{T}}V^{-1}D$:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx (D^{\mathrm{T}}V^{-1}D)^{-1}D^{\mathrm{T}}V^{-1}(\mathbf{y} - \boldsymbol{\pi}).$$

Considering the variation in $\hat{\boldsymbol{\beta}}$ as $\mathbf{y}$ varies, for fixed parameter values, we have, approximately

$$\mathrm{cov}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (D^{\mathrm{T}}V^{-1}D)^{-1}D^{\mathrm{T}}V^{-1}\,\mathrm{cov}(\mathbf{y} - \boldsymbol{\pi})V^{-1}D(D^{\mathrm{T}}V^{-1}D)^{-1}. \tag{18}$$

This simplifies to

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}) = (D^{\mathrm{T}}V^{-1}D)^{-1} \tag{19}$$

if $\mathrm{cov}(\mathbf{y} - \boldsymbol{\pi}) = V$, in other words, if the model variance assumption is correct, giving the familiar maximum-likelihood-type variance estimate based on the inverse information function.

The basic idea of the information-sandwich estimate is to substitute an empirical estimate of $\mathrm{cov}(\mathbf{y} - \boldsymbol{\pi})$ inside the 'sandwich' represented by (18). In general such an estimate may be obtained as

$$\hat{\mathrm{cov}}(\mathbf{y} - \boldsymbol{\pi}) = (\mathbf{y} - \hat{\boldsymbol{\pi}})(\mathbf{y} - \hat{\boldsymbol{\pi}})^{\mathrm{T}}$$

but this breaks down to a summation over block-diagonal components if we make the assumption of independence between 'primary sampling units'. In a longitudinal analysis, the latter are normally the individual subjects, leading to

$$D^{\mathrm{T}}V^{-1}\hat{\mathrm{cov}}(\mathbf{y}-\boldsymbol{\pi})V^{-1}D=\sum_i (D_i^{\mathrm{T}}V_i^{-1}(\mathbf{y}_i-\hat{\boldsymbol{\pi}}_i)(\mathbf{y}_i-\hat{\boldsymbol{\pi}}_i)^{\mathrm{T}}V_i^{-1}D_i)$$

where in practice we use estimated values of D and V based on $\hat{\boldsymbol{\beta}}$. Substitution of this expression in (18) leads immediately to the variance formula (11) given in the body of the tutorial.

ACKNOWLEDGEMENTS

This research was supported by a grant from Australia's National Health and Medical Research Council. The establishment and conduct of the cohort study was funded by the Victorian Health Promotion Foundation. We are also grateful to David Clayton, who both personally and through his short courses with Michael Hills (Perth, Western Australia, February 1996, and Cambridge, U.K., December 1996) significantly influenced our approach to the analysis of these data.

REFERENCES

1. Patton, G. C., Carlin, J. B., Coffey, C., Wolfe, R., Hibbert, M. E. and Bowes, G. 'The course of early smoking: a population based cohort study over three years', *Addiction*, **93**, 1251–1260 (1998).
2. Patton, G. C., Carlin, J. B., Coffey, C., Wolfe, R. and Bowes, G. 'Depression, anxiety and smoking initiation: a prospective study over three years', *American Journal of Public Health*, **88**, 1518–1522 (1998).
3. Stata Corp. *Stata Statistical Software: Release 5.0*, State Corporation, College Station, TX, 1997.
4. Ware, J. H. 'Linear modes for the analysis of longitudinal studies', *American Statistician*, **39**, 95–101 (1985).
5. Zeger, S. L. and Liang, K.-Y. 'Longitudinal data analysis for discrete and continuous outcomes', *Biometrics*, **42**, 121–130 (1986).
6. Neuhaus, J. M. 'Statistical methods for longitudinal and clustered designs with binary responses', *Statistical Methods in Medical Research*, **1**, 249–273 (1992).
7. Lipsitz, S. R., Kim, K. and Zhao, L. 'Analysis of repeated categorical data using generalized estimating equations', *Statistics in Medicine*, **13**, 1149–1163 (1994).
8. Hibbert, M, Hamill, M., Rosier, M., Caust, J., Patton, G. and Bowes, G. 'Computer administration of a school-based adolescent health survey', *Journal of Paediatrics and Child Health*, **32**, 372–377 (1996).
9. Lewis, G., Pelosi, A. J., Araya, R. and Dunn, G. 'Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers', *Psychological Medicine*, **22**, 465–486 (1992).
10. McCullagh, P. and Nelder, J. A. *Generalized Linear Models*, 2nd edn, Chapman and Hall, London, 1989.
11. Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.
12. Huber, P. J. 'The behaviour of maximum likelihood estimates under non-standard conditions', in *Fifth Berkeley Symposium in Mathematical Statistics and Probability*, University of California Press, Berkeley, CA, 1967, pp. 221–233.
13. White, H. 'Maximum likelihood estimation of misspecified models', *Econometrica*, **50**, 1–25 (1980).
14. Royall, R. M. 'Model robust confidence intervals using maximum likelihood estimators', *International Statistical Review*, **54**, 221–226 (1986).
15. Shah, B., Barnwell, B. and Bieler, G. *SUDAAN User's Manual, Release 7.0*, Research Triangle Institute, Research Triangle Park, NC, 1996.
16. Firth, D. 'Recent developments in quasi-likelihood methods', Proceedings of the ISI 49th Session, Firenze, 1993.
17. Liang, K.-Y. and Zeger, S. L. 'Longitudinal data analysis using generalized linear models', *Biometrika*, **73**, 13–22 (1986).
18. Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data*, Wiley, New York, 1987.

19. Binder, D. A. 'On the variances of asymptotically normal estimators from complex surveys', *International Statistical Review*, **51**, 279–292 (1983).
20. Flanders, W. D. and Greenland, S. 'Analytic methods for two-stage case-control studies and other stratified designs', *Statistics in Medicine*, **10**, 739–747 (1991).
21. Pickles, A., Dunn, G. and Vazquez-Barquero, J. L. 'Screening for stratification in two-phase ('two stage') epidemiological surveys', *Statistical Methods in Medical Research*, **4**, 73–89 (1995).
22. Clayton, D., Spiegelhalter, D., Dunn, G. and Pickles, A. 'Analysis of longitudinal binary data from multiphase sampling', *Journal of the Royal Statistical Society, Series B*, **60**, 71–87 (1998).
23. Little, R. J. A. and Schenker, N. 'Missing data', in Arminger, G., Clogg, C. C. and Sobel, M. E. (eds), *Handbook of Statistical Modeling for the Social and Behavioural Sciences*, Plenum Press, New York, 1995, pp. 39–75.
24. Kish, L. 'Weighting for unequal *P-i*', *Journal of Official Statistics*, **8**, 183–200 (1992).
25. Skinner, C. J., Holt, D. and Smith, T. M. F. *Analysis of Complex Surveys*, Wiley, Chichester, 1989.
26. Neuhaus, J. M., Kalbfleisch, J. D. and Hauck, W. W. 'A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data', *International Statistical Review*, **59**, 25–35 (1991).
27. Ware, J. H., Lipsitz, S. and Speizer, F. E. 'Issues in the analysis of repeated categorical outcomes', *Statistics in Medicine*, **7**, 95–107 (1988).
28. Berkey, C. S. and Laird, N. M. 'Nonlinear growth curve analysis: estimating the population parameters', *Annals of Human Biology*, **13**, 111–128 (1986).
29. Muthén, B. 'Latent variable modeling of growth with missing data and multilevel data', in Cuadras, C. M. and Rao, C. R. (eds.) *Multivariate Analysis: Future Directions* 2, Elsevier Science Publishers, Amsterdam, 1993, pp. 199–210.
30. Goldstein, H. *Multilevel Statistical Models*, 2nd edn, Arnold, London, 1995.
31. Bryk, A. S. and Raudenbush, S. W. *Hierarchical Linear Models: Applications and Data Analysis Methods*, Sage, Newbury Park, CA, 1992.
32. Dempster, A. P., Rubin, D. B. and Tsutakawa, R. K. 'Estimation in covariance components models', *Journal of the American Statistical Association*, **76**, 341–353 (1981).
33. Laird, N. and Ware, J. H. 'Random-effects models for longitudinal data', *Biometrics* **38**, 963–974 (1982).
34. Goldstein, H. and Rasbash, J. 'Improved approximations for multilevel models with binary responses', *Journal of the Royal Statistical Society, Series A*, **159**, 505–513 (1996).
35. Muthén, B. 'Analysis of longitudinal data using latent variable models with varying parameters', in *Best Methods for the Analysis of Change: Recent Advances, Unanswered Questions*, American Psychological Association, 1991, pp. 1–17.
36. Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980.
37. Collett, D. *Modelling Survival Data in Medical Research*, Chapman and Hall, London, 1994.
38. Clayton, D. and Hills, M. 'ssa10: analysis of follow-up studies with Stata 5.0', *Stata Technical Bulletin*, **40**, 27–39 (1997). Reprinted in *Stata Technical Bulletin Reprints*, **7**, 253–268 (1998).
39. Clayton, D. and Hills, M. *Statistical Models in Epidemiology*, Oxford University Press, Oxford, 1993.
40. Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. *Bayesian Data Analysis*, Chapman and Hall, London, 1995.
41. Robins, J. M., Rotnitzky, A. and Zhao, L. P. 'Analysis of semiparametric regression models for repeated outcomes in the presence of missing data', *Journal of the American Statistical Association*, **90**, 106–121 (1995).

NOTE ADDED IN PROOF

The details of Stata's survival time commands (Section 3.1) will need to be modified by users of Stata Version 6.0, released January 1999.

Part II
PROGNOSTIC/CLINICAL PREDICTION MODELS

2.1 Prognostic Variables

TUTORIAL IN BIOSTATISTICS

CATEGORIZING A PROGNOSTIC VARIABLE: REVIEW OF METHODS, CODE FOR EASY IMPLEMENTATION AND APPLICATIONS TO DECISION-MAKING ABOUT CANCER TREATMENTS

MADHU MAZUMDAR* AND JILL R. GLASSMAN

Division of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021, U.S.A.

SUMMARY

Categorizing prognostic variables is essential for their use in clinical decision-making. Often a single cutpoint that stratifies patients into high-risk and low-risk categories is sought. These categories may be used for making treatment recommendations, determining study eligibility, or to control for varying patient prognoses in the design of a clinical trial.

Methods used to categorize variables include: biological determination (most desirable but often unavailable); arbitrary selection of a cutpoint at the median value; graphical examination of the data for a threshold effect; and exploration of all observed values for the one which best separates the risk groups according to a chi-squared test. The last method, called the minimum p-value approach, involves multiple testing which inflates the type I error rates. Several methods for adjusting the inflated p-values have been proposed but remain infrequently used.

Exploratory methods for categorization and the minimum p-value approach with its various p-value corrections are reviewed, and code for their easy implementation is provided. The combined use of these methods is recommended, and demonstrated in the context of two cancer-related examples which highlight a variety of the issues involved in the categorization of prognostic variables.

1. INTRODUCTION

1.1. The need for prognostic variables

A *prognostic variable analysis* is a search for variables that predict the outcome of patients with differing characteristics, and is a large part of clinical research. Identifying variables that

* Correspondence to: Madhu Mazumdar, Division of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021, U.S.A. E-mail: mazumdar@biosta.mskcc.org

Contract/grant sponsor: Cancer Chemotherapy Program Project
Contract/grant number: CA 05826–35

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
 ISBN: 0-470-02365-1

characterize patients likely to have poor outcomes helps direct future research towards treatment refinements for targeted populations. In practice, prognostic variables are the tools clinicians use to determine which patients are candidates for moderate or extreme treatment measures. For example, tumour size is a variable that is predictive of a survival outcome for patients with breast cancer; patients with larger tumours have shorter survival times, and generally are recommended for more aggressive therapy.

1.2. The need for categorization

Changing a predictor variable from *continuous* to *categorical* form is a common part of many prognostic variable analyses, since categorization makes it easier for clinicians to use information about the relationship between an outcome and predictor variable in making treatment decisions. For a breast cancer patient, the risk of dying is an increasing function of the *continuous* variable tumour size. However, this information is not enough to direct patients towards different treatment options. A *cutpoint* is needed to separate patients into distinct risk groups having high or low probabilities of favourable outcomes. In fact, a tumour size cutpoint of 2 cm is used in part to classify breast cancer patients into such risk groups, which then suggest the recommended treatment path (for example, whether surgery alone is sufficient, or additional radiation therapy is necessary).[1]

Another reason for dichotomizing continuous prognostic variables such as age and blood protein levels is to set eligibility criteria for studies of new treatments. Many cancer studies exclude patients over the age of 60 – the cut-off believed to distinguish best between age-related treatment susceptibilities. Finally, continuous prognostic variables are categorized in order to 'stratify' patients in randomized clinical trials, so that roughly equal numbers of patients in differing risk groups are assigned to each treatment arm.

1.3. Methods of categorization

Many different methods are used for categorizing prognostic variables. Most common are visual inspections of simple scatter plots, designation of a cutpoint at an arbitrary percentile value, and systematic searches for the cutpoint associated with a minimum chi-squared p-value. Several authors have discussed the pros and cons of categorizing a continuous variable, and pitfalls encountered in a naive use of the methods for doing so.[2,3,5]

For searches in which cutpoints are systematically tested, Miller and Siegmund found a correction formula for the minimum p-value that accounts for having taken multiple, but not independent, looks at the data.[3] They also discussed the importance of examining effect sizes associated with potential cutpoints, such as relative risks, in addition to lone p-values. Lausen and Schumaker showed that this correction formula for the minimum p-value extends to the case of censored data;[5] Altman provided an easy formulation for the correction in order to further encourage its use.[2]

This paper reviews exploratory and systematic categorization methods whose combined use provides reassurance that the continuous prognostic variable has been reasonably categorized (Section 2). These methods are demonstrated in two comprehensive case studies. The first involves treatments for lymphoma and has a binary outcome variable (Section 3), and the second involves treatments for testicular cancer and has a censored outcome variable (Section 4). The computer code used to perform the analyses is provided in the Appendix.

2. METHODS

Ideally a cutpoint is suggested by theories of biological functioning, but this information is rarely available. Instead, observed data on the outcome and prognostic variable often are obtainable from a sample of patients, and can be explored in order to find empirically a cutpoint which appears to differentiate between high- and low-risk groups.

Once a continuous variable has been identified as being predictive of the outcome (generally by showing some association with the outcome in an appropriate regression setting), the first step is to examine a plot of the relationship between the two variables. It should appear that the underlying function is monotonic if a single cutpoint is sought; for a variable such as blood pressure, a single cutpoint cannot be used to divide patients into high- and low-risk groups, since values that are too high and too low both are associated with increased risk, as depicted in Figure 1, line b.[6]

Figure 1 lines a and c show monotonic functional relationships that might exist. The step function shown by line a is an ideal situation in which a threshold is apparent. More common are relationships that look like line c, where a cutpoint model appears not to provide the best fit. Nevertheless, a cutpoint often is needed for clinical decision-making. This is the case in the breast cancer example, where risk of death increases as tumour size increases, but it is necessary to fix some cut-off point below which patients are recommended for surgery.

A reasonable cutpoint can be identified if a data-based cutpoint exploration includes graphical as well as systematic analyses. This section describes the steps involved in performing selected graphical analyses, and integrating those results with the adjusted minimum p-value approach.

2.1. Exploratory analyses

Exploratory plots may reveal obvious thresholds that suggest potential cutpoints, or provide a range of values in which the search for a cutpoint should be performed. A list of plots for a variety of outcome types is suggested below:

(i) scatter plot;
(ii) grouped data plot;
(iii) lowess smoothed plot;
(iv) model-based predictive plot (censored outcome).

2.1.1. What to expect from the plots

For a binary outcome variable Y, a scatter plot over the observed values of the prognostic variable X ideally shows the degree to which the factor separates patients into risk groups. A cutpoint is revealed if the plot appears to be a step function (Figure 1 line a). Since such structure often is not apparent in typically noisy data, plotting means of Y (that is, proportions) for groups of X's is useful. Further smoothing the patterns in the proportions with a 'lowess' (locally weighted regression smoothed scatter plots) algorithm also helps uncover the underlying relationship. In the S-plus software package used in the case studies, lowess is a simple function which takes as input the observed (x, y) values, and the fraction f of data used for the smoothed estimate of y at each x point.[7]

When Y is a censored outcome variable such as survival time, a simple scatter plot of observed times over the values of the prognostic variable X is not informative since some of the y's are not fully observed. However, Kaplan–Meier[8] median survival estimates can be plotted for groups of

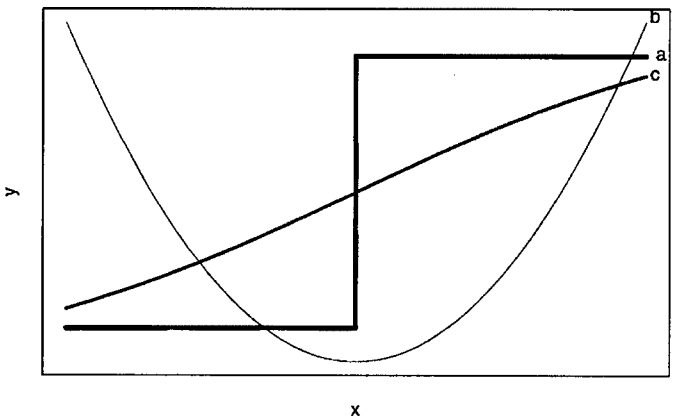


Figure 1. The generic outcome variable depicted in these plots is continuous. This may be because the underlying outcome is continuous, or was transformed into continuous form as a proportion, in the case of an underlying binary outcome, or as a model-based predicted survival time, in the case of a censored outcome. Line a: this is the ideal situation in which to use a single cutpoint model, since the underlying relationship between the outcome and prognostic variable is a step function. Line b: if the underlying relationship between the outcome and prognostic variable is not monotonic, a single cutpoint model cannot be used to divide patients into two distinct (high and low) outcome risk groups. Line c: commonly seen is an underlying monotonic relationship in which no cutpoint dividing patients into high and low outcome risk groups is apparent; reducing X to a dichotomous variable results in a loss of information, but is often necessary for practical treatment decision-making

x, when a sufficient proportion of the y's are fully observed in each group. Further smoothing can be accomplished using specific model-based predictive failure time plots, as described below.

2.1.2. Predictive failure time plots

Here, instead of plotting a summary of the y's for groups of x's, the time $\hat{y}$ at which a certain proportion q of patients remain alive is predicted for *each* x value. These predicted times can be derived from a Cox survival time model:

$$\hat{S}(y, x) = [\hat{S}_0(y)]^{\exp(x'\hat{\beta})} \tag{1}$$

where y is the observed survival time, x is the observed covariate value, $\hat{\beta}$ is the Cox regression coefficient, and $\hat{S}_0(y)$ is an estimate of the underlying survival function.[9] A particular qth percentile of the survival time distribution is then predicted by finding the value $\hat{y}$ which satisfies

$$\hat{S}_0(y) = (q)^{\exp(-x'\hat{\beta})}. \tag{2}$$

These predicted survival times are then plotted against the observed covariate values to look for monotonicity, a cutpoint, or a region in which a cutpoint search should be performed. Code for finding these predicted survival times is not straightforward, and is available from the author of reference 9.

2.2. Minimum *p*-value approach

Whether or not graphical examinations of the data suggest an obvious cutpoint, a systematic search further aids in the cutpoint selection process. In this approach, all observed values of the

prognostic variable, except a proportion of the extremes (see Section 2.3), are examined as candidates for the cutpoint. The value is chosen that best separates patient outcomes according to a maximum chi-squared statistic and minimum p-value, or a maximum relative risk.

For data with binary outcomes, at each observed value c in the interval of potential cutpoints, the following 2×2 contingency table is created:

	$X \leqslant c$	$X > c$
$Y = 0$	n_{11}	n_{12}
$Y = 1$	n_{21}	n_{22}

The value c which produces the maximum chi-squared statistic (or, equivalently, the corresponding minimum p-value) is selected as the cutpoint which best differentiates between outcome risk groups. Relative risks for each table also should be examined, since the chi-squared statistic is most sensitive to sample size.[3] In the case of a censored outcome, for each potential cutpoint c, a chi-squared statistic based on a logrank test,[10] corresponding p-value, and relative risk (generally based on a univariable Cox regression model) are computed in order to measure the strength of association between the dichotomized prognostic variable and the survival time endpoint.

A cutpoint found by testing each observed value in this systematic manner has been called 'optimal' by some users. Altman referred to the method as the minimum p-value approach in order to emphasize the problem of multiple testing. He demonstrated in a simulation that when this approach is used, the probability of obtaining a significant result (at the 5 percent level) from a logrank test,[10] when there is no actual relationship between the variables, is inflated to 40 per cent. Although sample size does not play a substantial role, the inflation increases as the interval of potential cutpoints is expanded.[2]

2.3. *p*-value adjustment formulae

Correction formulae have been derived in order to adjust for the inflation in the type I error rate that is associated with the minimum p-value approach. Miller and Siegmund derived p-value adjustment formula based on the asymptotic distribution for a maximum chi-squared statistic in the setting of a binary outcome variable.[3] The corrected p-value is calculated as

$$p_{\text{ms}} = \phi(z)\left(z - \frac{1}{z}\right)\log\left(\frac{\varepsilon_{\text{high}}(1 - \varepsilon_{\text{low}})}{(1 - \varepsilon_{\text{high}})\varepsilon_{\text{low}}}\right) + 4\frac{\phi(z)}{z} \tag{3}$$

where $p_{\min}$ is the observed minimum p-value, z is the $(1 - \frac{p_{\min}}{2})$th percentile of the standard normal distribution, and ε_{low} and $\varepsilon_{\text{high}}$ are, respectively, the proportion of observed values below the lowest and at or below the highest cutpoint value considered. This formula works best when there are a large number of potential cutpoints (> 50).[11]

No explicit guidelines for the choice of ε_{low} and $\varepsilon_{\text{high}}$ are provided in the literature. Miller and Siegmund's illustrative example, and Altman's simplified formulae (below) implicitly suggest eliminating the top and bottom 5 per cent or 10 per cent of the extreme values in the data. However, the choice of ε_{low} and $\varepsilon_{\text{high}}$ may be dictated in part by the nature of the particular data set being analysed (for example, see Section 4.4.3).

Altman *et al.* provided the following simplifications for formula (3). For $\varepsilon = \varepsilon_{high} = \varepsilon_{low} = 5$ per cent

$$p_{alt5} = -3{\cdot}13p_{min}(1 + 1{\cdot}65\ln(p_{min})) \tag{4}$$

and for $\varepsilon = \varepsilon_{high} = \varepsilon_{low} = 10$ percent

$$p_{alt10} = -1{\cdot}63p_{min}(1 + 2{\cdot}35\ln(p_{min})). \tag{5}$$

This approximation works well for small minimum p-values ($0{\cdot}0001 < p_{min} < 0{\cdot}1$) and is easy to apply.[2]

A standard Bonferroni correction would call for multiplying the minimum p-value by the number of cutpoints considered. This adjustment is appropriate only if the consecutive test statistics calculated are independent, which is not the case in a minimum p-value cutpoint search. A modified version of the Bonferroni correction was derived by Lausen and Schumaker, and considers the correlation between the test statistics for adjacent cutpoints. It is calculated by

$$p_{modbon} = p_{min} + \sum_{i=1}^{k-1} D(\varepsilon_i, \varepsilon_{i+1}) \tag{6}$$

where ε_i is the proportion of observed values at or below the ith cutpoint,

$$D(\varepsilon_i, \varepsilon_{i+1}) = \frac{\exp(-z^2/2)}{\pi}\left[a(\varepsilon_i, \varepsilon_{i+1}) - \left(\frac{z^2}{4} - 1\right)\left(\frac{a(\varepsilon_i, \varepsilon_{i+1})^3}{6}\right)\right]$$

and

$$a(\varepsilon_i, \varepsilon_{i+1}) = \sqrt{\left\{1 - \frac{\varepsilon_i(1-\varepsilon_{i+1})}{(1-\varepsilon_i)\varepsilon_{i+1}}\right\}}.$$

They also showed that p_{ms} and p_{modbon} apply when the outcome variable is censored. They recommended using the minimum of these two adjusted p-values, since the formulae tend to give conservative corrections.[5]

3. CASE STUDY 1: TREATMENT FOR UNRESPONSIVE LYMPHOMA

3.1. Treatment regimen and rationale for categorization

A standard approach to the treatment of patients with lymphoma that has not responded to conventional-dose chemotherapy is to administer it in high doses. Since high-dose (HD) chemotherapy is toxic to healthy blood cells as well as a patient's cancer cells, healthy blood cell precursors ('stem cells') are collected from a patient's blood prior to the HD chemotherapy; these are reinfused following the HD chemotherapy to help blood cells regrow. The greater the quantity of stem cells (SC) reinfused, the faster the patient recovers from the toxicity of the treatment. The number of stem cells collected from patients can in most cases be increased by the administration of a stem cell 'growth factor' in combination with conventional-dose chemotherapy. The progression of the key elements in this treatment regimen is enumerated below:

1. Conventional-dose chemotherapy + growth factor given (pre-collection regimen):
   (a) increases quantity of stem cells in blood;
   (b) reduces tumour burden.

2. X stem cells collected from blood → historical *cutpoint*, c_H used here:
   (a) If $X > c_H$ then go to step 3.
   (b) If $X < c_H$ then:
      (i) collect reserve SC from bone marrow (invasive) and go to step 3;
      (ii) otherwise, patient ineligible for HD portion of treatment.
3. High-dose chemotherapy given (primary curative regimen).
4. X stem cells reinfused.
5. Blood cells recovered after Y weeks.

In practice, clinicians have used their previous experience to define in advance some threshold dose c_H of stem cells necessary to ensure patients would have a 'rapid' recovery of their blood counts. In this case study, retrospective data collected from patients who went through steps 1–5 is used to derive systematically a new cutpoint c more appropriate for that population.[12]

3.2. Cutpoints found in literature

In the lymphoma literature, cutpoints defining threshold doses of stem cells needed for rapid recovery from HD chemotherapy range from 1·2 million to 5·0 million stem cells.[13–16] The most frequently reported cutpoint is 2·5 million cells, and Bensinger *et al.*[14] suggested that both 2·5 and 5·0 million cells be used to stratify the population into three risk groups. In all of the references reviewed, cutpoints were found simply by examining scatter plots for threshold effects.

3.3. Patients

The series analysed consists of 55 patients who were treated with an HD chemotherapy regimen at Memorial Sloan-Kettering Cancer Center between 1994 and 1997. All patients received the same conventional-dose chemotherapy and growth factor for their pre-collection regimen (step 1).

3.4. Outcome variable

Patients who have difficulty recovering blood counts are at a greater risk for infections, and require longer hospital stays, so that the risks and costs associated with their treatment are higher. The endpoint commonly used when identifying these patients is *time to recover* 20,000 *platelets* (PLT20K), in days since reinfusion of stem cells.[12–16] In the Memorial group of patients, this outcome is not censored since all patients achieved a platelet recovery.

Platelet recovery time is considered to be rapid when it occurs by the 14th day after the stem cells are reinfused.[12,13] Therefore, the binary outcome variable

$$Y = \mathrm{PLTC} = \begin{cases} 0 & \text{if PLT20K} < 14 \text{ days ('successful' platelet recovery)} \\ 1 & \text{if PLT20K} \geqslant 14 \text{ days (slow platelet recovery)} \end{cases}$$

is used in the following analyses. It would be natural to consider performing a cutpoint analysis with PLT20K as the prognostic variable and *survival* as the endpoint in order to confirm this dichotomization. However, this 14-day threshold is based not only on mortality but also on hospital costs and quality of life considerations. For the purposes of this example, the 14-day dichotomization of PLT20K is accepted

3.5. Prognostic variables

The variable that overwhelmingly predicts platelet recovery time in patients who receive HD chemotherapy is the quantity of stem cells reinfused (X = SC).[12,14] For the Memorial group of patients, univariable linear and logistic regression analyses indicated a positive association between SC and both PLT20K and PLTC (p-value = 0·04 and 0·02, respectively).

3.6. Methods and results

3.6.1. Exploratory methods

There is no known biological evidence that supports a particular threshold dose of stem cells.[13-16] A scatter plot of observed SC values for the group of patients who achieved successful platelet recoveries (PLTC = 0) and for those who did not (PLTC = 1) shows a fair amount of overlap between the two groups (Figure 2).

In order to get a better feel for the relationship between the quantity of stem cells reinfused and the likelihood of a rapid platelet recovery for these patients, observations were grouped into 14th percentiles according to the SC variable (14th percentiles appeared to best distil the signal from the noise in these data). For each group, the proportion of rapid platelet recoveries was calculated, and plotted against the group's mean SC quantity, as shown by the dots in Figure 3. While it looks like there was a higher chance for a rapid platelet recovery in patients who received greater SC quantities, it also appears that there may have been three platelet recovery risk groups.

To further tease out this pattern, a 'lowess' curve of the proportion of rapid platelet recoveries versus SC was superimposed in Figure 3. This was calculated using overlapping intervals of 70 per cent of the SC data.

Since the graphs show that monotonicity holds, a systematic search for a single cutpoint was appropriate, and was used to more definitively determine the cutpoint c for this population of patients. In the following section a single cutpoint is sought, and the possibility of a two-cutpoint model continues to emerge.

3.6.2. Minimum p-value approach

A minimum p-value analysis was performed in order to systematically divide patients into low-risk and high-risk platelet recovery groups. A wide interval of potential cutpoints was considered since the exploratory plots revealed no specific range in which to direct the search. Thus, for each observed SC value c between the 10th and 90th percentiles, a contingency table was created as shown in Section 2.2. A chi-squared statistic, corresponding p-value, and relative risk were computed to measure the relative strength of the association between PLTC and the dichotomized SC. For example, when c = 1·25 (the 10th percentile of SC), the table looked like

	SC $\leqslant$ 1·25	SC > 1·25
PLTC = 0	6	29
PLTC = 1	0	20

with corresponding χ^2 = 2·29, p = 0·13 and RR = 7·58 (continuity corrected). Results of these analyses for each c in the cutpoint interval are plotted in Figure 4; the code used to perform them is provided in the Appendix.

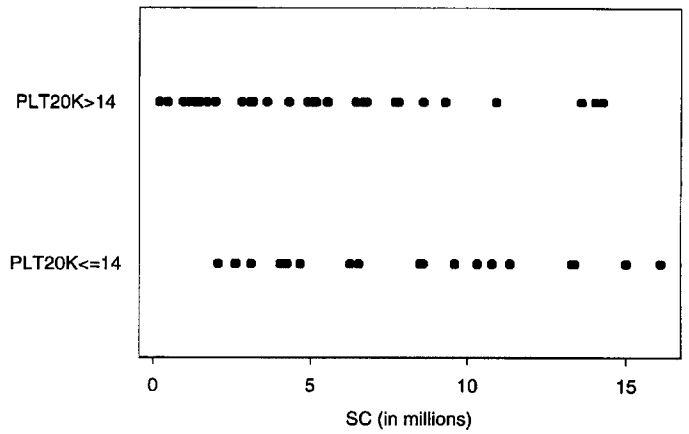


Figure 2. A scatter plot of the continuous prognostic variable SC (stem cell quantity) against the commonly accepted binary form of the platelet recovery time outcome variable is shown here. Thus the distribution of SC values for patients who achieved a quick platelet recovery, by day 14 (PLT20K > 14, or PLTC = 1), and for those who did not (PLT20K ⩽ 14, or PLTC = 0) are plotted together. An SC cutpoint would be revealed if this graph was a perfect step function

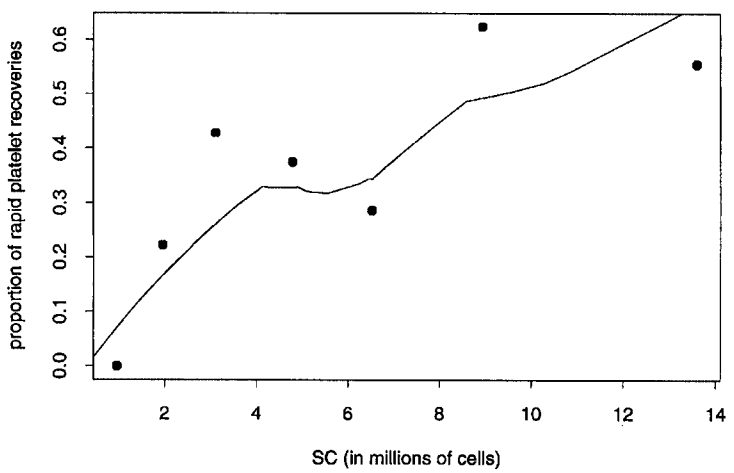


Figure 3. A plot of the prognostic variable SC against a continuous transformation of the binary platelet recovery time outcome (*proportion* of patients recovering rapidly, that is, by day 14) is shown here. In particular, the proportion of rapid platelet recoveries for each 14th percentile group of SC observations is plotted (14th percentiles appeared to best distil the signal from the noise in these data). The curve that is superimposed is a locally smoothed 'lowess' plot for which the mean value of the binary response variable PLTC was calculated for successively overlapping intervals of the factor SC (smoothing window size = 70 per cent; see lowess function description in reference 7)

For this series of patients, an SC cutpoint value of 2·0 was associated with a maximum chi-squared and (uncorrected) minimum p-value of 8·73 and 0·003, respectively. The relative risk also peaked at 2·0 million stem cells, but could not be reliably estimated because of the limited sample size (0·5 was added to all cells in each contingency table in order to obtain estimates in cases containing empty cells[17]).

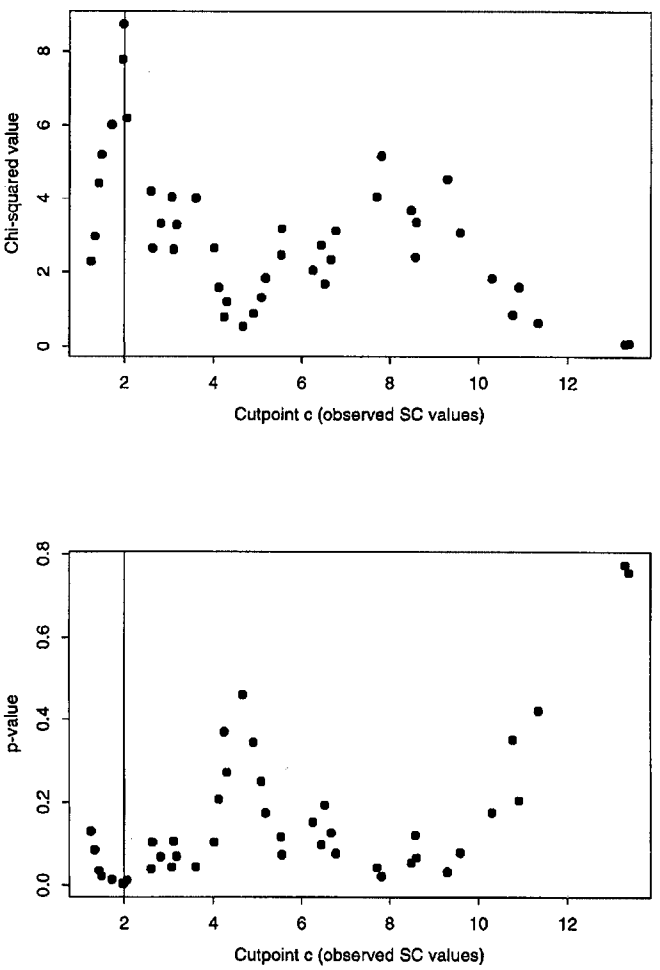


Figure 4. In the top figure, chi-squared values measuring the association between the PLTC outcome variable (rapid versus slow platelet recovery), and dichotomized SC are plotted as a function of the cutpoint c; in the bottom figure the corresponding chi-squared p-values are plotted as a function of c. While the maximum chi-squared value occurs at an SC level of 2 million cells, the plots suggest that there is a possible second SC cutpoint, defining a third risk group, at 8 million cells

3.6.3. Corrected p-values

In order to re-assess the actual prognostic significance of the SC variable dichotomized at 2·0 million cells, the corresponding minimum p-value of 0·003 was adjusted using the methods described in Section 2.3. The corrected p-values are shown below:

p_{min}	p_{ms}	p_{alt10}	p_{bon}	p_{modbon}
0·003	0·068	0·064	0·13	0·055

For this example, the p-value remains marginally significant according to any of the corrections except for the standard Bonferroni, which is expected to be overly conservative because it does not account for the correlations between the multiple tests. Based on the combined evidence of the exploratory and systematic analyses, 2·0 million cells appears to be a reasonable cutpoint for this patient group.

3.6.4. Additional observations

Figure 4 also shows a secondary peak at approximately 8 million stem cells, providing further evidence that there may be a second cutpoint. This would distinguish between low-risk patients, who receive more than 8 million SC, and medium-risk/high-risk patients. It should be noted that the methodology for finding a single cutpoint is based on a 2×2 contingency table model. Although this methodology may produce results that suggest two cutpoints, estimates of those would be found more appropriately by evaluating all potential cutpoint pairs (c_1, c_2) in a series of 2×3 contingency tables

	$SC < c_1$	$c_1 \leqslant SC < c_2$	$c_2 \leqslant SC$
PLTC = 0	n_{11}	n_{12}	n_{13}
PLTC = 1	n_{21}	n_{22}	n_{23}

No references were found that address this methodology, and the corresponding p-value adjustment formulae require further investigation.

4. CASE STUDY 2: TREATMENT FOR SEMINOMA

4.1. Background and rationale for categorization

The majority of patients with the testicular cancer called seminoma who are treated initially with chemotherapy are found to have a residual tumour mass. The management of these patients is controversial, and may involve surgery, radiotherapy, or close observation. If surgery is performed, the nature of the residual mass is established, and if found to be cancerous, it is removed, rendering the patient disease-free. However, 80–85 per cent of residual masses contain only non-cancerous tissue which do not require further therapy.[18] Therefore surgery is most justified for the subset of patients who are at greatest risk of having a cancerous residual mass. Identification of these high-risk patients using prognostic variables is essential to the treatment decision-making process.

The size of the residual mass on a post-chemotherapy computer tomography (CT) scan, measured as the largest diameter, is reported in the literature to predict for poor prognosis.[19,20] Puc *et al.* found this to be the only significant variable predictive of outcome (logrank p-value = 0·03).[21] They sought a CT diameter cutpoint to aid in identifying high-risk patients who were candidates for surgery. Their data set is used here to demonstrate how to find a cutpoint for the continuous CT diameter variable when the outcome variable is censored.

4.2. Patients

A total of 104 patients with advanced seminoma who were treated with various chemotherapy regimens at Memorial Sloan-Kettering Cancer Center from 1979 to 1992 were considered for this

retrospective analysis; 55 patients had post-chemotherapy surgery, and 49 patients were closely observed using CT scans following their chemotherapy. The median follow-up time was 4 years.

4.3. Outcome variable

Patients were designated as site failures or non-site failures. Site failure was defined as either the presence of cancerous tumour found at post-chemotherapy surgery, or clinical relapse at the assessed site during follow-up evaluation. Non-site failures had no cancerous tumour found at post-chemotherapy surgery, and no clinical relapse at the assessed site during follow-up evaluation. Failure-free survival time (time to site failure), calculated from the first day of chemotherapy treatment to the date of last follow-up evaluation, date of surgery, or date of relapse, was the outcome variable used in the analyses.

4.4. Methods and results

4.4.1. Exploratory methods

The frequency distribution of CT diameter is shown in Table I; the median CT diameter is 1·5 and ranges from 0·0 to 15·0. Figure 5 shows a scatter plot of the raw data, with time to site failure and non-site failure described by dots and stars, respectively. Most observed failure times occurred below a CT diameter of 5·0 cm. For patients with site failures, the range of CT diameters was 0·0–5·5, and for non-site failures the range was 0·0–15·0 (although only 4 per cent of the patients with non-site failures had CT diameters greater than 5·5). No CT diameter cut-off is suggested by these overlapping ranges. In fact, the general nature of the relationship between time to site-failure and CT diameter is not apparent from this scatter plot.

It was not possible to smooth these data by computing Kaplan–Meier median failure times for CT diameter groups, since an insufficient number of patients in the data set failed (10 out of 104). Therefore, a predictive failure time plot based on a Cox regression model was used to further investigate the functional relationship between CT diameter and failure-free survival time. The last column in Table I, labelled Ptime, lists the predicted failure-free survival times, which were computed using

$$\hat{S}_0(y) = (0{\cdot}9)^{\exp(-x'\hat{\beta})}. \tag{7}$$

The time at which 90 per cent of the patients remained failure free was selected for prediction due to the high proportion of censoring in the data. A line showing the predicted failure-free survival times (Ptimes) for each CT diameter value is superimposed on the scatter plot in Figure 5. This clearly shows that a higher CT diameter is associated with a greater risk of failure. The plateau at 41·9 months occurs simply because that is the maximum observed failure-free survival time, which is the maximum predicted under a Cox regression model. The curve also shows a steady decline in predicted failure-free survival times over a CT diameter interval of 2·5–5·0.

Further examination of the predicted failure-free survival times (Ptimes) in Table I reveals a separation of outcomes around CT diameters of 2·6 and 3·0. Based on empirical evidence, Motzer *et al.* selected a 3·0 cutpoint, recommending that patients with tumours below that size be closely observed without surgery.[20] As in case study 1, since the table and graphs show that monotonicity holds, a systematic minimum p-value analysis was appropriate, and was used to identify a specific cutpoint for this population.

Table I. Frequency distribution and predicted failure times for CT diameter

CT diameter	Frequency	Cum%	Ptime	CT diameter	Frequency	Cum%	Ptime
0·0	37	35·6	41·9	2·6	1	69·2	30·4
0·5	4	39·4	41·9	2·9	2	71·2	22·6
0·8	1	40·4	41·9	3·0	5	76·0	20·0
1·0	6	46·2	41·9	3·3	2	77·9	12·7
1·1	1	47·1	41·9	3·5	3	80·8	11·1
1·2	2	49·0	41·9	3·6	1	81·7	11·0
1·5	4	52·9	41·9	4·0	6	87·5	10·7
1·6	1	53·8	41·9	4·5	1	88·5	8·2
1·7	1	54·8	41·9	5·0	7	95·2	4·8
2·0	7	61·5	41·9	5·5	1	96·2	4·4
2·3	2	63·5	38·6	6·0	1	97·1	4·3
2·4	1	64·4	35·8	7·4	1	98·1	4·2
2·5	4	68·3	33·0	10·0	1	99·0	3·3
				15·0	1	100·0	2·7

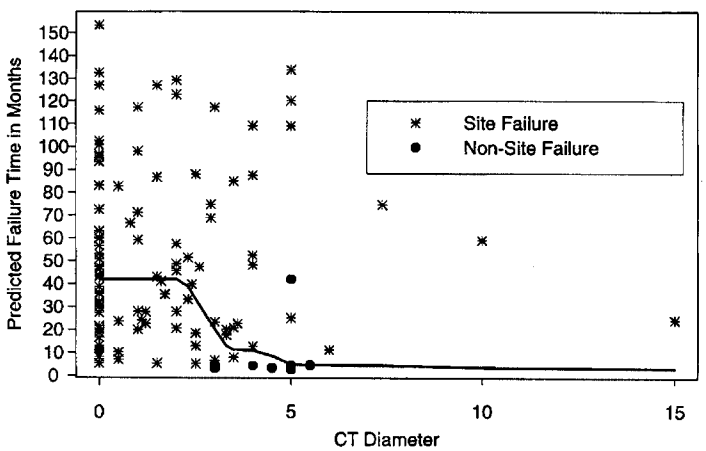


Figure 5. Cox predictive failure time plot for the prognostic variable CT diameter. The code is available from the author of reference 9

4.4.2. Minimum p-value approach

The failure time plot was used here to restrict the range of potential cutpoints to CT diameters between 2·0 and 5·0. CT diameter was categorized as CT $\leqslant c$ versus CT $> c$ for every observed c in the chosen range. A logrank test, corresponding p-value, and relative risk were computed for each cutpoint to measure the strength of association between the dichotomized CT diameter and the failure-free survival time. Results of these analyses are shown in Figure 6, and the code used to perform them is provided in the Appendix.

The recommended cutpoint was chosen by examining the plots of the relative risks and chi-squared values shown in Figure 6. The chi-squared values peaked at 16·0, corresponding to

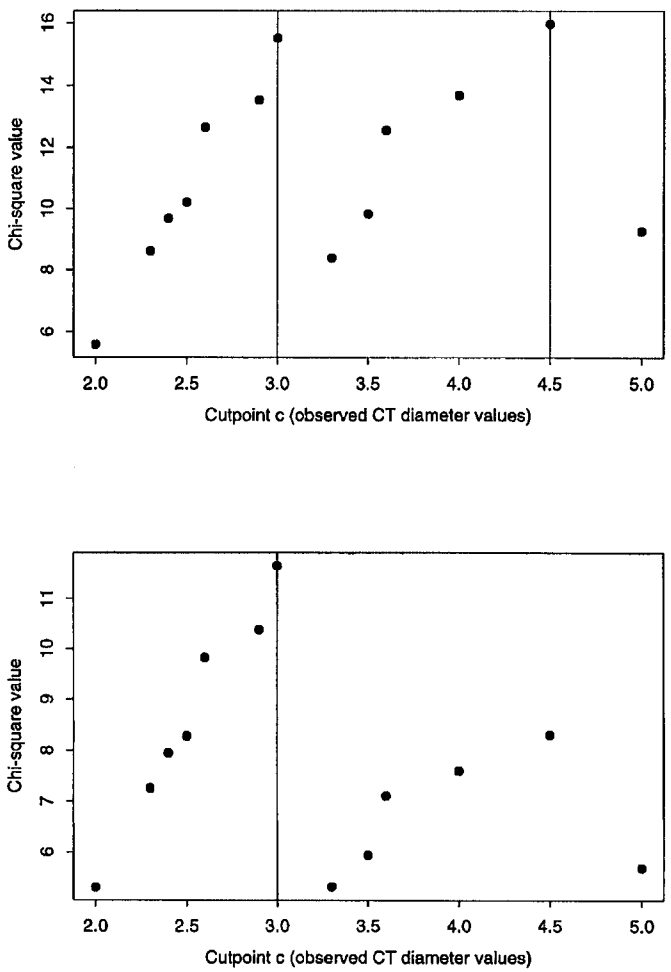


Figure 6. In the top figure, logrank chi-squared values measuring the association between failure-free survival time and dichotomized CT diameter are plotted as a function of the cutpoint c; in the bottom figure the relative risks are plotted as a function of c. While the maximum chi-squared value occurs at 4·5, a close second peak occurs at 3·0 where the relative risk also peaks, hence 3·0 was chosen as the cutpoint

a CT diameter of 4·5, with a close second peak at a chi-squared value of 15·5 and CT diameter of 3·0. Relative risks showed a clear single peak of 11·7 at a CT diameter of 3·0. Based on this and the evidence from the exploratory analyses, 3·0 was selected as the cutpoint.

4.4.3. Corrected p-values

Since the correction formulae apply only to the p-value corresponding to the maximum chi-squared statistic, the cutpoint search was redone to exclude all values at or above a CT diameter of 4·5. Therefore the value for ε_{high} to be used in formula (3) changed from 88·5 per cent to 87·5 per cent (see Table I). Because of the exploratory plot, the cutpoint search interval was restricted to

begin at a CT diameter of 2·0; in addition, 37/104 patients had an observed CT diameter of 0·0, so that ε_{low} for this case study was a relatively high 55 per cent.

The minimum p-value, corresponding to a CT diameter of 3·0, was then corrected using the appropriate methods described in Section 2.3. The following table shows the corrected p-values:

p_{min}	p_{ms}	p_{bon}	p_{modbon}
0·0001	0·0013	0·0009	0·0005

Although the p-value remains significant regardless of the correction used, it is interesting to note that for this example, p_{ms} is more conservative than even the standard Bonferroni correction. This is expected because of the small number of candidate cutpoints.[3,11]

5. FINAL SUMMARY

It is common practice to categorize a continuous prognostic variable for clinical use. This is done in order to set up practical eligibility criteria, stratification variables for clinical trials, or to guide clinicians and patients in their choice of therapy. This paper provided guidelines for applying a progression of exploratory categorization methods, together with the adjusted minimum p-value approach, in order to find the best cutpoint for the given data. Several p-value adjustment methods that account for the problem of multiple correlated testing, and re-assess the prognostic significance of the newly dichotomized variable, were discussed. The relative ease of applying these methods was demonstrated in two comprehensive case studies involving binary and censored outcomes, and programs for implementation were provided.

A thorough exploration of the data using percentile-grouped, smoothed, and predicted failure time plots helped to confirm the appropriateness of proceeding with the cutpoint analysis in case study 1, and suggested an interval on which to perform the search in case study 2. In case study 1, graphical exploration of the patterns in the chi-squared values over the entire range of cutpoints considered provided further insight into the possibility of a two-cutpoint model. In case study 2, explorations of the patterns in both the chi-squared values and relative risks over the entire range of candidate cutpoints ensured that a clinically relevant cutpoint was selected. Both of these observations would have been missed in a typical analysis focusing only on the minimum p-value.

Usually a prognostic variable is dichotomized in a univariable setting and then included in a multivariable model. While a univariable exploration of the most appropriate way for categorization is a necessary first step, it must be recognized that the cutpoint may depend on the levels of other independent prognostic variables. In reality, breast cancer treatment decisions are based on the 'TNM' (tumour size, nodal involvement, degree of metastasis) classification system, of which tumour size is only one factor.[1] For example, patients are classified in a particular risk category when tumour size is < 2 cm only if cancer is found in the lymph nodes; they are classified in that same risk category when tumour size is 2–5 cm if there is no lymph node involvement. Efforts have begun to incorporate cutpoint selection and p-value adjustments in a multivariable setting using classification and regression trees, but further evaluation of these methods is needed.[22,23]

In summary, the following steps are recommended for performing a comprehensive cutpoint analysis:

1. Identify continuous prognostic variable in appropriate regression setting.
2. Create exploratory plots (look for step pattern):
   (a) scatter plot of observed Y's versus X's;
   (b) plot of summary statistics of Y for grouped X's;
   (c) plot of lowess smoothed y curve over range of observed x's;
   (d) plot of predicted failure times $\hat{y}$ versus x.
3. Perform systematic search:
   (a) select candidate cutpoints of X from step 2;
   (b) dichotomize X at each cutpoint and compute association with Y, using:
      (i) chi-squared statistics and associated p-values;
      (ii) relative risks;
   (c) select cutpoint associated with min p-value, max chi-squared or relative risk;
   (d) re-assess prognostic significance of dichotomized X by adjusting p-value using:
      (i) Miller–Siegmund formula;
      (ii) Altman simplified adjustment formulae;
      (iii) modified Bonferroni adjustment (Lausen and Schumaker).

APPENDIX

Below is the code for the functions used to perform the cutpoint analyses presented in Sections 3 and 4. It is written in the S-plus programming language (version 3.3). A line begining with '#' indicates a comment. The code for each function is preceded by a list describing what it does, the input it requires, and the output it produces, to facilitate translation into other languages.

Function Name

MINP.

Function Performed

Evaluates potential cutpoints using chi-squared, p-value and relative risk criteria for data with binary outcomes (see 'minimum p-value approach', Section 2.2).

Input

x = vector of observed values of continuous prognostic factor;
ybin = vector of observed values of binary outcome variable;
xcutint = vector of potential cutpoints.

Output

An object (matrix) containing the following variables:
Cutpoint = sorted version of *xcutint*;
Chisquare = vector of chi-squared values; see table in Section 2.2;

pvalue = vector of *p*-values associated with above chi-squared values;
Relrisk = vector of relative risks for above table.

```
# MINP (x, ybin, xcutint)

function (x, ybin, xcutint)
{
    tmp1 ← sapply (sort(unique(xcutint)), function(x0, x, ybin)
    # sapply is a looping function that applies the function given as its
    # 2nd argument repeatedly to each element in the list given as its 1st argument
    {
            tmp ← chisq.test (1*(x < = x0), ybin)
            tabl ← table (1*(x > x0), ybin)
            rr ← ((tabl[1, 1] + 0·5)/((tabl[1, 1] + 0·5) + (tabl[2, 1] + 0·5)))/
            ((tabl[1, 2] + 0·5)/((tabl[1, 2] + 0·5) + (tabl[2, 2] + 0·5)))
    # x0 is the cutpoint being tested; its value is read from the list
    # given as the 1st argument to sapply
            c(x0, tmp$statistic, tmp$p.value, rr)
    # c = collect these items into a row matrix
    }
    , x, ybin)
    tmp1 ← data.frame(t(tmp1))
    # transpose to get a column instead of a row matrix
    names (tmp1) ← c("Cutpoint", "Chisquare", "pvalue", "Relrisk")
    tmp1
}
```

Function Name

MINPCENS.

Function Performed

Evaluates potential cutpoints using chi-squared, *p*-value and relative risk criteria for data with censored outcomes (see 'minimum *p*-value approach', Section 2.2).

Input

x = vector of observed values of continuous prognostic factor;
time = vector of observed values of outcome time to event;
status = vector of observed values of censoring indicator (0 means censored);
xcutint = vector of potential cutpoints.

Output

An object (matrix) containing the following variables:
Cutpoint = sorted version of *xcutint*;
Chisquare = vector of chi-squared values from log-rank tests; see Section 2.2;
pvalue = vector of *p*-values associated with above chi-squared values;
Relrisk = vector of relative risks based on univariable Cox regressions.

```
# MINPCENS
function(x, time, status, xcut)
{
        tmp1 ← sapply(unique(xcut), function(x0, x, time, status)
        {
            tmp ← surv.diff(time, status, 1 * (x < x0), rho = 0)
            tmp2 ← coxreg(time, status, 1 * (x < x0))
            c(x0, tmp$chisq, 1 - pchisq(tmp$chisq, 1), 1/exp(tmp2$coef))
        }
        , x, time, status)
        tmp1 ← data.frame(t(tmp1))
        #transpose to get a column instead of a row matrix
        names(tmp1) ← c("Cutpoint", "Chisquare", "pvalue", "Relrisk")
        tmp1
}
```

Function Name

PADJMS, PALT510, PMODBONF.

Function Performed

Computes the adjusted minimum *p*-value formulae derived by Miller and Siegmund, Altman, and Lausen and Schumaker (Section 2.2).

Input

Cutpoint = output vector from MINP;
pvalue = output vector from MINP;
epsi.high = proportion of observed values of factor x that are at or below the highest cutpoint value tested;
epsi.low = proportion of observed values of factor x that are below the lowest cutpoint value tested;
x = vector of observed values of continuous prognostic factor.

Output

Cut.point = (scalar) the *Cutpoint* associated with the minimum *pvalue*;
$p - min$ = (scalar) the minimum *pvalue*;
epsi.high = the value input for *epsi.high*;
epsi.low = the value output for *epsi.low*;
*pms.palt*5, *palt*10, *pmodbon f* = the adjusted minimum *p*-values.

```
#PADJMS(Cutpoint, pvalue, epsi.high, epsi.low)
function(Cutpoint, pvalue, epsi.high, epsi.low)
{
        pmin ← min(pvalue)
        Cut.point ← Cutpoint[pvalue = = min (pvalue)]
        z ← qnorm(1 − pmin/2)
        f.z. ← dnorm(z)
        pacor ← f.z. * (z − 1/z) * log((epsi.high * (1 − epsi.low))/((
        1 − epsi.high) * epsi.low)) + (4 * f.z)/z
```

```
        pval ← c(Cut.point, round(pmin, 6), epsi.high, epsi.low,
        round (pacor, 6))
        names(pval) ← c("Cut.point", "p-min", "epsi.high", "epsi.low",
        "pms")
        pval
}

#PALT510(Cutpoint, pvalue)
function(Cutpoint, pvalue)
{
                pmin ← min(pvalue)
                Cut.point ← Cutpoint [pvalue = = min(pvalue)]
                pcor10 ← − 1·63 * pmin * (1 + 2·35 * log(pmin))
                pcor5 ← − 3·13 * pmin * (1 + 1·65 * log(pmin))
                pval ← c(Cut.point, round(pmin, 6), round(pcor5, 6), round(
                        pcor10, 6))
                names(pval) ← c("Cut.point", "p-min", "palt5", "palt10")
                pval
}

#PMODBONF(x, Cutpoint, pvalue)

function(x, Cutpoint, pvalue)
{
        pmin ← min(pvalue)
        Cut.point ← Cutpoint[pvalue = = min(pvalue)]
        z ← qnorm(1 − pmin/2)
        f.z. ← dnorm(z)
        n ← length(x)
        dsum ← 0
        for(i in 1:(length(Cutpoint) − 1)) {
                l ← length(x[x <= Cutpoint [i]])
                ll ← length(x[x <= Cutpoint [i + 1]])
                t ← sqrt(1 − (l * (n − ll))/(n − 1) * ll))
                d ← sqrt(2/3·14) * f.z * (t − (((z^2)/4 − 1) * t^3)/6
                        )
                dsum ← dsum + d
        }
        pmodbonf ← pmin + dsum
        pval ← c(Cut.point, round(pmin, 6), round(pbonf, 6))
        name(pval) ← c("Cut.point", "p-min", "pmodbonf")
        pval
}
```

ACKNOWLEDGEMENTS

The authors thank their colleague Dr. E.S. Venkatraman for generously sharing his experience on the topic, Ms. T. Polyak for help in writing the programs, and Dr. K.S. Panageas for thoughtful comments which helped distil key ideas in the paper. They also thank Dr. C. Moskowitz, Dr. E. Hedrick and Dr. R. Motzer for providing access to the data used in the examples and advice about the clinically relevant issues. Finally, the detailed comments provided by the reviewer helped tremendously to improve the clarity of the manuscript. This work was supported in part by the Cancer Chemotherapy Program Project (CA 05826-35).

REFERENCES

1. Gilewski, T., Norton, L. 'Breast cancer' *in* Kelley, W. N. (ed.), *Disorders of Oncology and Hematology*, Lippincott-Raven Publishers, Philadelphia, 1997, Chapter 221.
2. Altman, D. G., Lausen, B. and Sauerbrei, W. 'Dangers of using "optimal" cutpoints in the evaluation of prognostic factors', *Journal of the National Cancer Institute*, **86**, (11), 829–835 (1994).
3. Miller, R. and Siegmund, D. 'Maximally selected chi square statistics', *Biometrics*, **38**, 1011–1016 (1982).
4. Lausen, B. and Schumacher, M. 'Maximally selected rank statistics', *Biometrics*, **48**, 73–85 (1992).
5. Lausen, B. and Schumacher, M. 'Evaluating the effect of optimized cutoff values in the assessment of prognostic factors', *Computational Statistics & Data Analysis*, **21**, 307–326 (1996).
6. Hansson, L., Zanchetti, A., Carruthers, S. G., Dahlof, B., Elmfeldt, D., Julius, S., Menard, J., Rahn, K. H., Wedel, H. and Westerling, S. 'Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principle results of the Hypertension Optimal Treatment (HOT) trial', *Lancet*, **351**, 1755–1762 (1998).
7. *S-PLUS for Windows Reference Manual*, Vol. 1, version 3.1, Statistical Sciences, Inc., Seattle, WA, March, 1993.
8. Kaplan, E. L. and Meier, P. 'Nonparametric estimation from incomplete observation.' *Journal of the American Statistical Association*, **53**, 457–481 (1958).
9. Heller, G. and Simonoff, J. S. 'Prediction in censored survival data: a comparison of the proportional hazards and linear regression models', *Biometrics*, **48**, 101–115 (1992).
10. Mantel, N. 'Evaluation of survival data and two new rank order statistics arising in its consideration', *Cancer Chemotherapy Reports*, **50**, 163–170 (1966).
11. Hilsenbeck, S. G. and Clark, G. M. 'Practical p-value adjustment for optimally selected cutpoints', *Statistics in Medicine*, **15**, 103–112 (1996).
12. Moskowitz, C., Glassman, J., Wuest, D., Maslak, P., Reich, L., Gucciardo, A., Coady-Lyons, N., Zelenetz, A. and Nimer, S. 'Factors affecting mobilization of peripheral blood progenitor cells in patients with lymphoma', *Clinical Cancer Research*, **4**, 311–316 (1998).
13. Haas, R., Mohle, R., Fruhauf, S., Goldschmidt, H., Witt, B., Flentje, M., Wannenmacher, M. and Hunstein, W. 'Patient characteristics associated with successful mobilizing and autografting of peripheral blood progenitor cells in malignant lymphoma', *Blood*, **83**, (12), 3787–3794 (1994).
14. Bensinger, W., Appelbaum, F., Rowley, S., Storb, R., Sanders, J., Lilleby, K., Gooley, T., Demirer, T., Schiffman, K., Weaver, C., Clift, R., Chauncey, T., Klarnet, J., Montgomery, P., Petersdorf, S., Weiden, P., Witherspoon, R. and Buckner, C. 'Factors that influence collection and engraftment of autologous peripheral-blood stem cells', *Journal of Clinical Oncology*, **13**, (10), 2547–2555 (1995).
15. Bender, J., Bik To, L., Williams, S. and Schwartzberg, L. 'Defining a therapeutic dose of peripheral blood stem cells', *Journal of Haematotherapy*, **1**, 329–341 (1992).
16. Hohaus, S., Goldschmidt, H., Ehrhardt, R. and Haas, R. 'Successful autografting following myeloablative conditioning therapy with blood stem cells mobilized by chemotherapy plus rhG-CSF', *Experimental Haematology*, **21**, 508–514 (1993).
17. Agresti, A. *Categorical Data Analysis*, Wiley, New York, 1990.
18. Herr, H. W., Scheinfeld, J., Puc, H. S., Heelan, R., Bajorin, D. F., Mencel, P., Bosl, G. J. and Motzer, R. J. 'Surgery for a post-chemotherapy residual mass in seminoma', *Journal of Urology*, **157**, (3), 860–862 (1997).
19. Jones, B. M., Newlands, E. S., Begent, R. H., Rustin, G. J., Bagshawe, K. D., Johnson, A. G. and Reynolds, K. W. 'The role of abdominal surgery in the treatment of advanced testicular germ cell tumors', *British Journal of Surgery*, **69**, (1), 4–6 (1982).
20. Motzer, R., Bosl, G., Heelan, R., Fair, W., Whitmore, W., Sogani, P., Herr, H., and Morse, M. 'Residual mass: an indication for further therapy in patients with advanced seminoma following systemic chemotherapy', *Journal of Clinical Oncology*, **5**, 1064–1070 (1987).
21. Puc, H., Heelan, R., Mazumdar, M., Herr, H., Scheinfeld, J., Vlamis, V., Bajorin, D., Bosl, G., Mencel, P. and Motzer, R. 'Management of residual mass in advanced seminoma: results and recommendations from the Memorial Sloan-Kettering Cancer Center', *Journal of Clinical Oncology*, **14**, 454–460 (1996).
22. Lausen, B., Sauerbrei, W. and Schumacher, M. 'Classification and regression trees (CART) used for the exploration of prognostic factors measured on different scales', in Dirschedl P and Ostermann R. (ed.), *Computational Statistics*, Physica-Verlag Heidelberg, Germany, 1994.
23. Lausen, B., Kersting, M. and Schoch, G. 'The regression tree method and its application in nutritional epidemiology', *Informatik, Biometric und Epidemiologie in medizin und biologie*, **28**, (1), 1–13 (1997).

2.2 Prognostic/Clinical Prediction Models

DEVELOPMENT OF HEALTH RISK APPRAISAL FUNCTIONS IN THE PRESENCE OF MULTIPLE INDICATORS : THE FRAMINGHAM STUDY NURSING HOME INSTITUTIONALIZATION MODEL

RALPH B. D'AGOSTINO, ALBERT J. BELANGER

Mathematics Department, Boston University, 111 Cummington Street, Boston, MA 02215, U.S.A.

ELIZABETH W. MARKSON

Gerontology Center, Boston University, 53 Bay State Road, Boston MA 02215, U.S.A.

AND

MAGGIE KELLY-HAYES AND PHILIP A. WOLF

Preventive Medicine and Epidemiology, Department of Medicine, 80 East Concord Street, Boston, MA 02118, U.S.A.

SUMMARY

A health risk appraisal function is a mathematical model designed to estimate the risk or probability of a person's mortality or morbidity for various diseases based upon risk factors such as age, medical history and smoking behaviour. The Framingham Study has contributed substantially to the development and use of these for endpoints such as mortality and incidence of coronary heart disease and other cardiovascular diseases. This paper discusses a methodology for the development of health risk appraisal functions when the number of potential risk factors is large and illustrates it with sex specific functions for nursing home institutionalization. The methodology involves grouping variables substantively into sets, applying principal component factor analysis and variable clustering to obtain substantively meaningful composite scores, ranking these in order of substantive importance, and then entering these with a hierarchical ordering into a Cox proportional hazard regression.

1. INTRODUCTION

A health risk appraisal function (or risk profile function) is a mathematical model that has as input an individual's risk factors, or risk profile, and produces an assessment of the risk that the individual will develop a particular event within a stated time. For example, the Framingham stroke function estimates the probability that a stroke-free individual will develop a stroke within 10 years given the subject's stroke risk profile that consists of sex, age, systolic blood pressure, smoking behaviour, diabetes status, previous cardiovascular disease history, a diagnosis of atrial fibrillation, left ventricular hypertrophy and the use of anti-hypertension medication.[1,2] One can also use this function to estimate the relative risk of developing stroke where one compares the individual to the 'average' person of the same age and sex. Some risk appraisal functions are more elaborate. For example, the Carter Center's Health Risk Appraisal computer program produces

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
 ISBN: 0-470-02365-1

probability estimates for 41 specific causes of mortality and also an estimate of the potential number of years of life lost associated with the risk factor profile of the individual.[3]

The Framingham Study[4,5] has contributed substantially to the area of health risk appraisal functions, having developed functions for the assessment of risk of incidence of and mortality from various cardiovascular diseases such as myocardial infarction, coronary heart disease, congestive heart failure and stroke.[1,2,6-10] Evaluations of these Framingham functions have shown that they are valid and reliable.[11-16] The Carter Center's coronary heart disease mortality health risk appraisal function, which was evaluated to be the most reliable among a number of health risk appraisal functions,[15] was also a Framingham function supplied by the present authors to the Carter Center.

These past Framingham functions have been simple regression models such as logistic regressions,[6-8] Cox regressions[1,2] or accelerated failure models,[9,10] characterized by the presence of a small number of risk factors, such as systolic blood pressure, total serum cholesterol and smoking, identified as strong predictors of cardiovascular disease over the 40 years history of the Framingham Study.

As the Framingham Study cohort has aged, there has been an increasing number of the participants institutionalized in nursing homes. The study investigators desired to produce health risk appraisal functions suitable for estimating the probability or risk of institutionalization in a nursing home within a time period of up to, say, 5 years. Such a function would contain predictors of institutionalization considered simultaneously, and could make important contributions in reducing, delaying or planning placement in at least four ways:

1. One could alert health professionals to pay special attention to the development and progression of risk factors in the elderly that may threaten their capacity to remain at home.
2. Programme planners in a variety of settings could use the risk appraisal functions to devise or modify programmes to meet the specific needs of those elders most at risk.
3. One could alert family members and the elders themselves to these risk factors, some of which may be preventable or reduced. Knowledge of one's own risk profile, or that of a family member, would provide a prospective assessment of the likelihood of nursing home placement that would markedly assist advance planning.
4. Identification of the relative contribution of risk factors could have utility in planning policies for long term care financing. Elders with certain risk profiles represent different potential levels of costs that one can forecast. This would be information useful to the individuals, their families, as well as for public policy.

There is a literature relating social and cognitive factors in the elderly (age $\geqslant 65$ years) to institutionalization. Factors such as age, sex, living alone, social isolation, distance from relatives, marital status, income, race and dementia have been identified as associated with it.[17-21] Others have investigated functional ability and have noted, for example, that disability scores are strongly predictive.[22] Further, medical factors such as worsening health have also been identified.[23] These studies tended to consider factors in a univariate manner, have often dealt only with nursing home patients with non-institutionalized controls, and did not attempt to quantify the simultaneous effects of these variables. The Framingham Study provides a unique opportunity to consider the multivariate aspects. Thus a project was initiated to develop health risk appraisal functions for nursing home institutionalization that would consider prospectively and simultaneously the contributions of social, cognitive, functional and medical factors.

In contrast to previous Framingham Study investigations, a small number of appropriate risk factors was not delineated. Rather, there was proposed a large number of candidate variables or *indicators* that quantify the social, cognitive, functional and medical domains. Most of these

Table I. Steps in the development of a health risk appraisal function in the presence of multiple indicators

Population and model selection
1. At risk population and outcome event selected (select at risk population, that is, the sample for the analysis and select the endpoint, that is, the dependent variable for the analysis).
2. Select mathematical model (candidate models are logistic regression, Cox regression, parametric survival models where length of follow-up and validity of assumptions of the models are important to consider).
Data preparation and data reduction
1. Group data into substantive sets (Delphi-like sessions involving, at least in part, experts in substantive field).
2. Perform principal components factor analysis and variable cluster analysis (employ empirical data reduction techniques to produce composite scores. Other techniques such as battery reduction are possible).
Production of health risk appraisal function
Perform hierarchical regression (or other methods that will ensure substantively important variables are entered and chance fluctuations are minimized).
Further refinements
Test for interactions, examine models on subsets, verify assumptions of the models, examine residuals and examine calibration and discrimination properties.
Validation
Employ cross-validation, jack-knife and/or bootstrap. Evaluate model on a test data set and/or on a new data set.

indicators have been collected carefully at Framingham. The development of the health risk appraisal functions with these indicators presented a number of new and interesting features not common to the analyses usually employed on the Framingham Study data, but which the authors have dealt with extensively in the context of producing predictive functions in the medical decision arena.[24-27]

The following presents details of the methodologic issues in the development of these new health risk appraisal functions. The objective is not to produce final validated functions but rather to focus on the methodologic aspects of their development. We provide advice and comments on the practical issues, using for illustration the development of the nursing home institutionalization model. Table I contains a flow chart of the steps in the development. Continued reference to this should aid in understanding the proposed method.

2. POPULATION AND MODEL SELECTION

2.1. At risk population and outcome event

The Framingham Study is a major prospective epidemiologic study begun in 1948–52 with 5209 participants (2336 males and 2873 females) aged 30 to 62.[4,5] Since then, these subjects have been re-examined approximately every two years. At examination 17 (1981–83), the exam of the first systematic collection of functional assessment measures, the survivors were aged 61 to 93. Of these, there were 2104 (842 men and 1262 women) not institutionalized and evaluated at examination 17. During the subsequent six years, 126 (31 men and 95 women) were admitted to nursing homes or other long term care facilities. For developing the risk appraisal functions the 2104 constitute the at risk population and the 126 the cases or events.

From examination 17, there were data collected on 105 variables on the 2104 subjects that related to social, cognitive, functional and medical aspects. Details on these variables appear below.

2.1.1. Practical issues and advice

In developing health risk appraisal functions it is important to define clearly what is the population at risk. In the present case, the at risk population consists of all those evaluated at examination 17 and not institutionalized at that time. On any examination in Framingham, there is usually about an 85 per cent response rate. For examination 17,440 individuals alive and known not to reside in a nursing home did not attend. It was desired not to impute data on their cognitive and functional variables, so they have been excluded in the development of the health risk appraisal function. Some of these subjects did appear for examination 18. Such subjects were added subsequently to the data set and used in the analysis for a shorter time period (that is, three years). Their inclusion did not affect any of the results. Also, the 440 did not differ from the 2104 on age, sex, subsequent institutionalization or other major events such as death rates and cardiovascular diseases.

In general, investigators should have concern that exclusion of subjects may create biases. At a minimum, one should undertake comparisons of their baseline characteristics such as age, sex and morbidity with the characteristics of those included in the model development.

A second important step is to decide upon the endpoint, the dependent variable in the health risk appraisal function. For the present function the endpoint is the first institutionalization. It is possible that a person admitted to a nursing home was later discharged. Such an individual is considered as having experienced an event. Another endpoint might have been institutionalization for at least some time period, or until death. In the present case this variation had no effect on the final function, but in other situations it may have.

Most of the published Framingham Study health risk appraisal functions deal with first events (such as first episode of cardiovascular disease) where the at risk population are free of the endpoint at the baseline evaluation. The endpoint event, however, could simply be the occurrence of a particular event, such as a stroke, without reference to whether it is the first such event. In such a case, we strongly recommend inclusion as one of the independent variables in the analysis the existence of a previous event. In general, we have found that functions that have first events as the endpoint and the at risk population as those who are disease free at baseline are easier to interpret than other functions.

2.2. Mathematical model

The aim of the model development is to find an 'optimal' subset of the 105 variables and relate these, by means of a Cox regression, to the survival function for institutionalization (that is, the probability of not being institutionalized by a time t), given symbolically by

$$S(t|\mathbf{X}) = S_0(t)^{\exp(\boldsymbol{\beta}'\mathbf{X})}. \tag{1}$$

Here $S(t|\mathbf{X})$ is the conditional probability of not being institutionalized by time t $(0 < t < 5)$ given the vector of variables $\mathbf{X}$, where $\mathbf{X}$ represents the vector of variables that 'best' relate to institutionalization, $\boldsymbol{\beta}$ is the vector of regression coefficients and $S_0(t)$ is the survival function for $\mathbf{X} = \mathbf{0}$. The probability of institutionalization by time t is $F(t|\mathbf{X}) = 1 - S(t|\mathbf{X})$.

2.2.1. Practical issues and advice

Many Framingham health risk appraisal functions with a short follow-up, such as up to 4 or 5 years, are based on the logistic regression model.[3,6-8] The present study used the Cox proportional hazard regression function. The main advantages of the Cox regression is that it takes into account the timing of the event, thus potentially increasing the precision of its estimates, and it can deal with censored observations such as deaths within the 5 year period. Of course, to insure the validity of the Cox analysis, one has to verify the proportional hazard assumption. If this assumption is not verified, one could employ an accelerated failure model, such as that employed by Anderson *et al.*[10]

3. DATA PREPARATION AND DATA REDUCTION

The data analysis for the nursing home institutionalization function consisted of two phases. First, there was the data preparation and reduction phase and second there was the development of the risk appraisal functions separately for each sex. The objective of the first phase was to insure the availability of stable, reliable variables for inclusion in the model. We now discuss the first phase.

3.1. Grouping data into substantive sets

As noted already, there were a large number of variables or indicators, many measuring the same underlying *dimension* or underlying *structure*. For example, there were a number of variables that measured functional ability. The term *indicator* here signifies that the observed variables measure these underlying dimensions. Some of these may be latent and not directly measurable, such as, for example, depression. We undertook data preparation and data reduction to produce composite variables that summarize these dimensions. The data preparation consisted of *grouping the variables on substantive grounds*. Of the 105 variables, we generated 16 substantive sets. They were:

1. *Demographic* (including age, sex and marital status)
2. *Medication use* (including cardiac medications, diuretics, diabetes medication, sleeping pills, antidepressive medication, etc.)
3. *Perception of health* (physician perception)
4. *Activities of daily living* (*ADL*) such as the Katz ADL scales[28]
5. *Functional performance* variables such as personal care and grooming[29]
6. *Gait* variables consisting of ever a hip fracture and a new hip fracture
7. *Sensory* variables such as visual impairment variables
8. *Neurological* variables including presence of a stroke, Folstein mini-mental examination questions,[30] physician evaluation of dementia and other measures of dementia
9. *Depression* variables related to sleep disturbances
10. *Respiratory* variables such as dyspnoea and bronchospasm variables
11. *Cardiac* variables such as hypertension, angina, myocardial infarction, and congestive heart failure
12. *Vascular* variables including measures of peripheral vascular disease
13. *Obesity* containing one variable, body mass index
14. *Cancer* containing presence of cancer
15. *Diabetes* containing a diagnosis of diabetes or taking medication for diabetes
16. *Alcohol use* measuring the amount of alcohol per week.

The study researchers generated this groupings of the variables employing a Delphi approach.[31]

3.1.1. Practical issues and advice

At this stage one must apply great care and much energy. For example, a stepwise statistical analysis, no matter how carefully done, is not a suitable replacement. Those researchers involved in the substantive aspects of the project, such as the geriatric physicians in the present project, need to understand the seriousness of this step. A formal detailed Delphi session is one approach, but may be too extensive. In the present project one geriatric physician took the lead by first developing sets of variables for discussion. These were then discussed in detail with the other investigators, including other physicians, nurses, social scientists and statisticians, many expert in geriatrics. The process resulted in the above grouping.

Ideally, the substantive sets generated should be exclusive, that is, no variable is included in two different sets. This, however, is not always sensible or appropriate. For example, in the present study the respiratory and congestive heart failure symptoms of 'dyspnoea on exertion' and 'dyspnoea increased in the past two years' were included both in the respiratory (10) and cardiac (11) sets.

3.2. Principal component factor analysis and cluster scores

The next phase in the analysis was to generate composite functions of the variables within the substantive sets by empirical techniques. For those sets that consisted of only one variable, such as the obesity set (13) which consisted of only body mass index, we took that variable as the measure or composite function of the set and its corresponding underlying dimension. For those substantive sets that contained two or more variables, we performed a principal components factor analysis within each set, followed by the determination of composite scores with use of variable cluster analysis.[32] These final composite scores were cluster scores that we considered to quantify the underlying dimensions of the substantive sets. The methods employed here are described in Cureton and D'Agostino (Chapters 5, 6, 8, 9, 12 and 14).[32]

For these analyses, we combined data from the males and females. Further, because of the differences in scales of the variables within substantive sets, we first standardized all variables and then entered them into the analyses. For the principal component factor analysis, we retained all components with eigenvalue greater than unity. We then employed both varimax and Promax rotations to obtain the simple structure solutions. We considered all variables with factor loadings 0·4 or larger in the appropriate factor matrices to define the underlying factor and we took these variables as a *cluster* of variables for the factor. The two rotation procedures produced similar results. When there were differences, we took the Promax solution as the preferred one. This exercise did produce meaningful subsets of variables within the substantive sets already produced above from clinical consideration. For example, the cardiac set (set 11 above) consisted of 17 variables and in the factor analysis it produced 4 factors: a congestive heart failure symptoms factor; a hypertension factor; an angina factor, and a myocardial infarction factor. From the 105 variables there were 16 substantive sets. From these, there were 30 factors or clusters of variables generated by the factor analysis.

We next generated cluster scores for each of the clusters of variables that contained more than one variable.[32] These cluster scores and those single variables from the substantive sets with single variables were the predictor or independent variables used in the next phase, the development of the risk appraisal functions. *In the following we refer to both types of variables (cluster scores or single variables) as cluster variables or simply as variables.*

3.2.1. Practical issues and advice

The principal component version of factor analysis is a useful technique for data reduction (or dimension reduction) in the sense that one can use it to identify the dimension of the data and those variables that correlate highly with one another. It accomplishes this in a two-stage process, as described in Chapters 5, 6, 8, 9 and 12 of Cureton and D'Agostino.[32] First, one transforms the original variables into a set of uncorrelated linear functions of the variables where these transformed variables 'explain' a substantial proportion of the original variables. The number of linear functions is usually much less than the original number of variables and one can think of this number of functions as the dimension of the data. Second, one then transforms these linear functions to a new set of linear functions where the objective of this last transformation is to produce composite functions interpretable by examination of the similarity of the original variables which have large coefficients in them (see below for more detail on this).

More formally, we summarize principal components as follows. Assume we have n variables, $X_1, \ldots, X_n$, all of which are standardized, and we desire to generate m composite variables of the form

$$Y_j = w_1 X_1 + \cdots + w_n X_n \quad \text{for } j = 1, \ldots, m \tag{2}$$

where $m < n$ and w_i are selected to 'explain' the maximum possible variance. That is, in (2), Y_1 will have the largest possible variance (λ_1) subject to the restriction $w_1^2 + \cdots + w_n^2 = 1$ (that is, the weights normalized), Y_2 will be uncorrelated with Y_1 and have the next largest possible variance (λ_2 with $\lambda_1 > \lambda_2$) etc., until we obtain m such uncorrelated composite variables all of which have weights normalized and variances

$$\lambda_1 > \cdots > \lambda_m . \tag{3}$$

The percentage of the variance of the original n variables explained by the m composite variables is

$$100(\lambda_1 + \cdots + \lambda_m)/n. \tag{4}$$

We call the m composite scores in this context component scores. There are many standard statistical software packages (for example, the SAS Procedure PRINCOMP) one can use to perform the principal components.[33] One can also employ the SAS Procedure FACTOR.

The first step in principal components is to determine m, the number of components to retain. One popular rule is to let m equal the number of components with variances λ_i greater than one. There are other possibilities (see Cureton and D'Agostino, Chapter 12[32]). The second step is to obtain interpretable composite components. The usual procedure for this step is to produce the initial component matrix **A** (sometimes referred to as initial factor **F**) and *rotate* it. **A** (or **F**) is an n by m matrix containing the correlations of the original n variables to the m components (or m factors). The objective of the rotation is to produce weights in the composite variables of (2) that are large on a small number of the original variables and close to zero on the other variables. This is often called simple structure (see Cureton and D'Agostino, Chapter 6[32]).

In the Framingham Heart Study a 10 question depression scale had been administered on examination 18 where the responses were no or yes to 10 questions such as 'life is an effort', 'I am happy', etc. For the subjects of this study there were three principal components with variances greater than unity. They were 3·357, 1·290 and 1·022 for a percentage variance explained equal to 56·69 per cent. The initial component matrix **A**, the rotated factor matrix and the weights w_i for the composite functions of (2) appear in Table II, for analysis variables were scaled so that high scores indicated depression.

The rotation matrix is the Promax Reference matrix. It is an oblique rotation. See Cureton and D'Agostino, Chapters 6, 8 and 9 for details.[32] The last three columns contain weights w_i for the composite functions.

Table II. Principal component analysis of depression variables

	Initial matrix A			Rotation matrix			Weights w_i		
EFFORT	0·60	0·15	0·41	0·07	0·60	0·06	0·03	0·42	0·04
RESTLESS	0·39	0·07	0·55	−0·08	0·64	−0·12	−0·04	0·45	−0·08
DEPRESS	0·77	−0·13	−0·10	0·62	0·13	0·06	0·27	0·08	0·05
HAPPY	0·70	−0·23	−0·06	0·61	0·12	−0·06	0·26	0·07	−0·04
LONELY	0·64	−0·23	−0·21	0·65	−0·03	−0·00	0·28	−0·03	0·00
UNFRIEND	0·35	0·67	−0·33	0·04	−0·06	0·80	0·02	−0·04	0·59
ENJOYLIFE	0·52	−0·27	−0·27	0·63	−0·13	−0·03	0·28	−0·10	−0·02
FELTSAD	0·71	−0·22	−0·20	0·69	0·00	0·01	0·30	−0·01	0·01
DISLIKED	0·34	0·72	−0·22	−0·06	0·04	0·79	−0·02	0·03	0·58
GETGOING	0·58	0·20	0·47	0·01	0·66	0·07	−0·01	0·46	0·05

As can be seen from the columns of weights in Table II, even after the rotation, all the variables tend to be included in all component composite scores. Variable cluster analysis techniques attempt to group variables in non-overlapping sets, or equivalently set some of the weights in the composite functions equal to zero so that each variable is in at most one composite function. A an INTUITIVE CLUSTER ANALYSIS an examination of the rotation matrix would select variables with large loadings (in the present example loadings greater than 0·4) to define the cluster and then set, for all the other variables not in the cluster, the corresponding weights in the composite functions equal to zero. From Table II this would select the DEPRESS, HAPPY, LONELY, ENJOYLIF and FELTSAD variables to form one cluster. EFFORT, RESTLESS and GETGOING a second one and UNFRIEND and DISLIKED a third cluster. In addition to setting some weights equal to zero, often the variables in a cluster are given equal weights. For the above example the first cluster composite function then is

$$C_1 = \text{DEPRESS} + \text{HAPPY} + \text{LONELY} + \text{ENJOYLIF} + \text{FELTSAD}.$$

Note, in the present institutionalization study, the variables were standardized. This was due to the fact that not all variables in the original substantive sets had the same scale. If they have the same scale and have approximately the same variance, the sum of the original unstandardized variables is usually the preferred cluster score.

For developing the institutionalization health risk appraisal functions, we first employed this intuitive clustering. It produced, however, the same results as more formal variable clustering analyses, as described in Cureton and D'Agostino, Chapter 14.[32] SAS Macros for these formal procedures have been generated by Ralph B. D'Agostino, Joseph Massaro, Kimerly Dukes and Zhini Zhang and are available upon request from one of the present authors (RBD). The SAS Procedure VARCLUS also clusters variables.

Another dimension reduction technique is to select m variables in each substantive set that reproduced, as much as possible, the variance of the original n variables. This is called *battery reduction* and is described in Cureton and D'Agostino, Chapter 12.[32] A SAS Macro for this is also available upon request from one of the authors (RBD). We did not apply this method in the present project.

Two other topics merit mentioning. The first relates to the use of Likert and dichotomous scale variable data in principal components. If the variables employed in the analysis are dichotomous, there often are components or factors formed that relate to the frequency of a response and not necessarily to some underlying dimension. For example, if one administers a questionnaire to

a sample with a high frequency of males and a large number of stroke cases, one will find these two variables related by virtue of the fact that both will receive a high frequency of responses even though they do not measure an underlying substantive component. This 'inappropriate' component is often called a 'difficulty factor' because of the research where it first was studied.[34-36] The researcher should be aware of this and should not allow for consideration of these components as real components.

The second issue that merits mentioning relates to the linearity of the variables in principal components. This technique implicitly assumes linearity. There are methods that automatically solve for variable transformations and may deal better with the input variables.[37, 38]

Lastly, for further discussion of principal components the reader is referred to the books by Jackson[39] and Jolliffe.[40]

4. PRODUCTION OF THE HEALTH RISK APPRAISAL FUNCTIONS

4.1. Hierarchical regression

The final objective was to produce 'health risk appraisal functions', one for each sex. We selected the Cox regression model of equation (1) as the functional form of the model. To guard against capitalization on chance fluctuation and to ensure meaningful results, we arranged the cluster variables in sets (or levels) in order of their substantive importance and then entered them into a *hierarchical* regression. Recall, here cluster variables refers to the cluster scores and also to the single variables for those sets of variables in Section 3.1 that contained only one variable. We used $p = 0{\cdot}05$ for the determination of statistical significance of a cluster variable. If the p-value of a cluster variable was above 0·05 ($p > 0{\cdot}05$) upon entry, or was above 0·05 in the presence of other cluster variables, we removed it from the model. The ordering of the sets of cluster variables into levels for entry into the multivariate model was as follows: (1) age; (2) marital status; (3) activity clusters; (4) functional clusters; (5) hip fracture variables (new or old fracture); (6) dementia cluster; (7) stroke diagnosis; (8) personal care clusters; (9) body mass index; (10) cardiac clusters including the congestive heart failure cluster; (11) alcohol use; (12) diabetes; (13) physician perception of health; (14) antidepressant medications prescribed; (15) depressed sleep; (16) sensory problems; (17) respiratory difficulties, and (18) cancer. We tested for statistical significance of a cluster variable in a level or set of cluster variables only after we had tested for significance all those cluster variables from all previous levels. Removal could have occured anywhere in the process.

4.1.1. Practical issues and advice

The major objectives of this stage are to ensure that we allow entry of the important substantive cluster variables into the final model if they are significant and that we minimize the influence of chance fluctuations. Arranging the cluster variables (cluster scores or single variables from those substantive sets with single variables) into levels or sets by a hierarchical ordering of substantive importance and then testing separately that the cluster variables within each level of the hierarchical set are statistically significant is an efficient means of achieving these goals. The method employed in the present study considered each level of the hierarchy after previous levels were examined. At each level we entered all the variables in the level and then we performed a backward elimination procedure to eliminate all variables in that level or previous levels with significance above 0·05. When only variables with significance of $p \leqslant 0{\cdot}05$ remained, we entered the variables from the next level. This continued until we had examined all levels of cluster variables.

Table III. Health risk appraisal function for nursing home institutionalization

	Cox regression coefficients for multivariate model											
	Men (31 events/842 subjects)						Women (95 events/1262 subjects)					
	b	s.e.	*p*	RR	95%	CI	*b*	s.e.	*p*	RR	95%	CI
Demographic												
Age	0·12	0·03	0·0001	1·13	1·07	1·19	0·10	0·02	0·0001	1·10	1·07	1·14
Married	−0·30	0·43	0·48	0·74	0·32	1·72	−0·73	0·27	0·008	0·48	0·28	0·82
Activity level												
Low intensity activity	0·29	0·13	0·028	1·15*	1·02	1·31						
Gait												
Hip fracture – new							1·76	0·41	0·0001	5·80	2·61	12·91
Neurological												
Dementia	0·39	0·17	0·023	1·22*	1·03	1·44	0·45	0·09	0·0001	1·25*	1·15	1·36
Cardiac												
CHF symptoms	0·49	0·16	0·002	1·28*	1·09	1·49	0·25	0·12	0·038	1·13*	1·01	1·27
Others												
BMI							−0·08	0·02	0·001	0·92	0·88	0·97
Diabetes							0·80	0·28	0·004	2·23	1·30	3·85

* relative risk for an increment of one-half unit – all others for a change of one unit
Variables defined as follows:

Age in years

Married: 0 if no, 1 if yes

Low intensity activity = $1{\cdot}34X_1 + 1{\cdot}33X_2 + 1{\cdot}08X_3$
where score for X_i ($i = 1, 2, 3$) is from self-report data;
1 if need help from other person and/or equipment, 0 otherwise
X_1: need help to go from bed to chair
X_2: need help to walk across a small room
X_3: need help to walk up and down stairs to second floor

Gait: 1 if hip fracture within last two years, 0 otherwise

Dementia = $0{\cdot}10Y_1 + 0{\cdot}54Y_2$
where
Y_1 = 30 − Folstein Mini Mental Score
Y_2: physician assessment of subject's mental status with scores 1 if normal, 2 if normal but has a physical impairment, 3 if possibly demented and 4 if dementia present

CHF (congestive heart failure) symptoms = $0{\cdot}36Z_1 + 0{\cdot}79Z_2 + 0{\cdot}73Z_3 + 0{\cdot}87Z_4 + 0{\cdot}65Z_5 + 0{\cdot}85Z_6 + 0{\cdot}22Z_7$
where score for $Z_i (i = 1, \ldots, 6)$ is 1 if symptoms present, 0 otherwise
Z_1: dyspnoea increased in last 2 years
Z_2: recent orthopnoea
Z_3: old orthopnoea
Z_4: paroxyamal nocturnal dyspnoea
Z_5: congestive heart failure diagnosed in last 2 years
Z_6: clinical diagnostic impression of congestive heart failure
Z_7: dyspnoea on exertion with scores 0 if no, 1 if yes with vigorous exercise, 2 if yes with rapid walking, 3 if yes with any slight exercise

BMI (body mass index) = weight in kilograms/(height in metres)^2
Diabetes: (diagnosed or taking drugs) 0 if no, 1 if yes

Table IV. Underlying survival function for sex-specific models

	$S_0(t)$	
	Men	Women
1 year	0·9976	0·9967
2 years	0·9932	0·9866
3 years	0·9892	0·9789
4 years	0·9853	0·9724
5 years	0·9805	0·9629
β_0	8·3471	4·7285

A forward stepwise procedure is an alternative. It is useful if we have a large number of variables at each level.

As stated above, the method is to test for significance of variables at a particular level only after one has entered into the model all variables in the previous levels and have tested them for significance. A variation of this is to find the significant variables within each level, separately and independently of the other levels. Then, one could examine in a hierarchical manner only the significant variables of the different levels.

Even though in the above we present methods for selecting and removing variables based only on statistical significance, it is important to review all variables not in the final model for substantive importance. There may be important variables which should be included even if they do not attain statistical significance. See the article by Spiegelhalter on the hazards of deleting 'insignificant' variables.[41]

Lastly, this hierarchical regression step in the health risk appraisal model development is basically the employment of regression methods to obtain a prognostic model. Standard techniques exist for these and one should incorporate them as appropriate.[42, 43]

4.2. Final models

The final sex specific models appear in Table III. The model for males contains five variables (two single variables and three cluster variables) and the model for females contains seven variables (four single variables and three cluster variables).

Table III also contains relative risk and their 0·95 confidence intervals to indicate the impact of each variable. One variable for the males (marital status) did not attain statistical significance. We retained it in the model because we considered it of substantive importance and its relative risk is in line with the literature.[17-19]

One can compute probabilities for institutionalization for each sex using the following formula:

$$S(t|\mathbf{X}) = S_0(t)^{\exp(\boldsymbol{\beta}'\mathbf{X} - \beta_0)}. \tag{5}$$

Here $S(t|\mathbf{X})$ is the conditional probability of not being institutionalized by time $t (0 < t < 5)$ given the vector of variables $\mathbf{X}$ for a particular individual. We obtain $S_0(t)$ of equation (5) from Table IV for $t = 1\,(1)\,5$, $\boldsymbol{\beta}$ in $\exp(\boldsymbol{\beta}'\mathbf{X} - \beta_0)$ is the vector of coefficients of the variables given in Table III and β_0 is a correction term for the model and is in Table IV. Usually the Cox regression model appears as in formula (1) where the $\mathbf{X}$ represents variables subtracted from their mean values. We can use formula (5) directly with the actual data of a subject. The term β_0 makes the adjustment for the deviations from the means. Note that there is a separate model for each sex.

5. FURTHER REFINEMENTS

We investigated interactions by considering the collection of variables formed by cross products and entering them as sets of variables in the regressions with adjustment for multiple testing. No interaction was identified as significant. We generated further regressions for subsets (married, unmarried, different age groups). None differed from the full model, another indication that interactions terms are unnecessary. We tested proportionality of the variables and examined residuals. In no cases did we detect any problems. Lastly, we examined calibration and discrimination properties of the models.

6. DISCUSSION

As outlined in the Section 1, there are a number of uses for the health risk appraisal functions for institutionalization. These can be used by health professionals, families and the elderly themselves to assess the risk of institutionalization. These individuals can be made sensitive to risk factors that may relate to institutionalization. One can design further programmes to address the major variables that relate to institutionalization.

The methodology employed involves beginning with a large number of variables, first arranging them by substantive considerations, and second employing principal component analysis and cluster analysis for dimension reduction and creating cluster scores for entry into a regression model. One then orders these cluster variables in terms of their substantive importance and enters them into the model in a backwards elimination, or forward stepwise fashion, or in a forced manner. This methodology should prove useful for generating health risk appraisal models with other outcomes where there are a large number of variables that compete for entry into the model.

In this paper, for illustration of the methods, we present only sex specific models. Also we are developing a model combining the sexes. There are a number of significant interactions of sex and the other cluster variables making the model more complicated than the sex specific models.

The above material focused on model development. There is also the problem of validation. A traditional procedure for dealing with this is to employ the concept of a developmental set. Here, one develops the original model on the developmental set and then tests it on a test data set. Other approaches are to use techniques such as jack-knifing or bootstrapping. Because of the small number of events (31 out of 837 males and 95 out of 1254 females or 3·7 per cent and 7·5 per cent, respectively), especially for the males, validation is an important issue. We are performing presently a bootstrap analysis. Previously we computed a number of models for specific subsets of subjects. These subset models did not differ for the model on the full data. This does give assurance of validation. Finally, the real test of validity is to use the model on an independent data set or an independent set of subjects. This activity is also underway.

ACKNOWLEDGEMENTS

This research was funded by National Heart, Lung and Blood Institute grant 1-RO1-HL-40423-05 and by AARP Andrus Fund.

REFERENCES

1. Wolf, P. A., D'Agostino, R. B., Belanger, A. J. and Kannel, W. B. 'Probability of stroke: a risk profile from the Framingham Study', *Stroke*, **22**, 312–318 (1991).
2. D'Agostino, R. B., Wolf, P. A., Belanger, A. J. and Kannel, W. B. 'Stroke profile: adjustment for antihypertensive medication: the Framingham Study', *Stroke*, **25**, 40–43 (1994).
3. Hutchins, E. B. *Healthier People: The Carter Center of Emory University Health Risk Appraisal Program, Vol. 4*, Emory University, Atlanta, Georgia, 1988.

4. Dawber, T. R. *The Framingham Study: the Framingham Study: the Epidemiology of Atherosclerotic Disease*, Harvard University Press, Cambridge, Massachusetts, 1980.
5. D'Agostino, R. B. and Kannel, W. B. 'Epidemiological background and design: the Framingham Study', *Proceedings of the American Statistical Association Sesquicentennial Invited Paper Sessions*, American Statistical Association, Alexandria, Virginia, pp. 707–718, 1989.
6. Truett, J., Cornfield, J. and Kannel, W. B. 'A multivariate analysis of the risk of coronary heart disease in Framingham', *Journal of Chronic Diseases*, **20**, 511–524 (1967).
7. Gordon, T., Kannel, W. B. and Halperin, M. 'Predictability of coronary heart disease', *Journal of Chronic Diseases*, **32** (3), 427–440 (1979).
8. Kannel, W. B. and McGee, D. 'Composite Scoring – methods and predictive validity: insights from the Framingham study', *Health Services Research*, **22**, 499–535 (1987).
9. Anderson, K. M., Wilson, P. W. F., Odell, P. M. and Kannel, W. B. 'An updated coronary risk profile: a statement for health professionals', *Circulation*, **83**, 356–362 (1991).
10. Anderson, K. M., Odell, P. M., Wilson, P. W. F. and Kannel, W. B. 'Cardiovascular disease risk profiles', *American Heart Journal*, **121**, 293–298 (1991).
11. 'Pooling Project Research Group. Relationship of blood pressure, serum cholesterol, smoking habit, relative weight and ECG abnormalities to incidence of major coronary events: final report of the Pooling Project', *Journal of Chronic Diseases*, **31**, 201–306 (1978).
12. Leaverton, P. *et al.* 'Representativeness of the Framingham risk model for coronary heart disease mortality: a comparison with a national cohort study', *Journal of Chronic Diseases*, **40**, 775–784 (1987).
13. Gordon, T. and Kannel, W. B. 'Multiple risk functions for predicting coronary heart disease: the concept, accuracy and application', *American Heart Journal*, **112**, 1031–1039 (1982)
14. Katz, D. and Foxman, B. 'How well do prediction equations predict? Using receiver operating characteristic curves and accuracy curves to compare validity and generalizability', *Epidemiology*, **4**, 319–326 (1993).
15. Smith, W. S., McKinlay, S. M. and McKinlay, J. B. 'The reliability of health risk appraisals: a field trial of four instruments', *American Journal of Public Health*, **79**, 1603–1607 (1989).
16. Smith, K. W., McKinlay, S. J. and McKinlay, J. B. 'The validity of health risk appraisals for coronary heart disease: results from a randomized field trial', *American Journal of Public Health*, **81**, 466–470 (1991).
17. Cohen, M. A., Tell, E. J. and Wallack, S. S. 'The risk factors of nursing home entry among residents of six continuing care retirement communities', *Journal of Gerontology/Social Sciences*, **43**, S15–S21 (1988).
18. Shapiro, E. and Tate, R. 'Who really is at risk of institutionalization?', *The Gerontologist*, **28**, 237–245 (1988).
19. Burr, J. A. 'Race/sex comparisons of elderly living arrangements: factors influencing the institutionalization of the unmarried', *Research on Aging*, **12**, 507–530 (1990).
20. Branch, L. G. and Jette, A. M. 'A prospective study of long term care institutionalization among the aged', *American Journal of Public Health*, **72**, 1373–1379 (1982).
21. Branch, L. B. 'Relative risk rates of nonmedical predictors of institutional care among elderly persons', *Comprehensive Therapy* **10**, 33–40 (1984).
22. Manton, K. G. 'A longitudinal study of functional change and mortality in the United States', *Journal of Gerontology/Social Sciences*, **43**, S153–S161 (1988).
23. Hing, E. 'Use of nursing home by the elderly: preliminary data from the 1985 National Nursing Home Survey', *Advanced Data*, No 135. Hyattsville, MD, Public Health Services, 14 May 1987.
24. Pozen, M. P., D'Agostino, R. B., Mitchell, Rosenfeld, M., Guglielmino, J. T., Schwartz, M. L., Teebagy, N., Valentine, J. M. and Hood, W. B. 'The usefulness of a predictive instrument for reducing inappropriate admissions to the coronary-care unit', *Annals of Internal Medicine*, **92**, 238–242 (1980).
25. D'Agostino, R. B. and Pozen, M. 'The logistic function as an aid in the detection of acute coronary heart disease in emergency patients (a cast study) ', *Statistics in Medicine*, **1**, 41–48 (1982).
26. Pozen, M. W., D'Agostino, R. B., Selker, H. P., Sytkowski, P. A. and Hood, W. B. 'A predictive instrument to improve coronary-care unit admission practices in acute ischemic heart disease: a prospective multicenter clinical trial', *New England Journal of Medicine*, **319**, 1273–1278 (1984).
27. Selker, H. P., Griffith, J. L. and D'Agostino, R. B. 'A tool for judging coronary unit admission appropriateness, valid for both real-time and retrospective use; a time-insensitive predictive instrument (TIPI) for acute cardiac ischemia: a multicenter study', *Medical Care*, **29**, 610–627 (1991).
28. Katz, S. and Downs, T. D. 'Progress in the development of the index of ADL', *The Gerontologist*, **10**, 20–30 (1970).

29. Rosow, I. and Breslau, N. 'A Guttman health scale for the aged', *Journal of Geronotology*, **21**, 556–559 (1966).
30. Folstein, M. F., Folstein, S. and McHugh, P. R. 'Mini mental state exam: A practical method for grading the cognitive state of patients for clinicians', *Journal of Psychiatric Research*, **12**, 192–198 (1975).
31. Miholland, A. V., Wheeler, S. C. and Heieck, J. J. 'Medical assessment by a Delphi group opinion technique', *New England Journal of Medicine*, **298**, 1272–1275 (1973).
32. Cureton, E. E. and D'Agostino, R. B. *Factor Analysis, An Applied Approach*, Erlbaum Publishers, New Jersey, 1983.
33. SAS Institute Inc. *SAS User's Guide Statistics, Version 6.03 Edition*, SAS Institute Inc., Cary, North Carolina, 1990.
34. Campbell, D. T. and Fiske, D. W. 'Convergent and discriminant validation by the multitrait multimethod matrix', *Psychological Bulletin*, **56**, 81–105 (1959).
35. Flamer, S. 'Assessment of the multitrait multimethod matrix validity of likert scales vs confirmation analysis', *Mulivariate Behavioral Research*, **18**, 275–308 (1983).
36. Matell, M. S. and Jacoby, J. 'Is there an optimal number of alternatives for likert scale items? Study I: reliability and validity' *Educational and Psychological Measurements*, **31**, 657–674 (1971).
37. Kuhfeld, W. F. 'The PRINQUAL procedure', in *SAS/STAT User's Guide*, Vol. 2, Chapter 34, SAS Institute, Inc., Cary, NC, 1990, pp. 1265–1323.
38. LeBlanc, M. and Tibshirani, R. 'Adaptive principal surfaces', *Journal of the American Statistical Association*, **89**, 53–64 (1994).
39. Jackson, J. E. *A User's Guide to Principal Components*, Wiley, New York, 1991.
40. Jollife, I. T. *Principal Component Analysis*, Springer–Verlag, New York, 1986.
41. Spiegelhalter, D. J. 'Probabilistic prediction in patient management', *Statistics in Medicine*, **5**, 421–433 (1986).
42. Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B. and Rosati, R. A. 'Regression modeling strategies for improved prognostic prediction', *Statistics in Medicine*, **3**, 143–152 (1984).
43. Harrell, F. E., Lee, K. L., Matcher, D. B. and Reichert, T. A. 'Regression models for prognostic prediction: advantages, problems, and suggested solutions', *Cancer Treatment Reports*, **69**, 1071–1077 (1985).

TUTORIAL IN BIOSTATISTICS

MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS

FRANK E. HARRELL Jr., KERRY L. LEE AND DANIEL B. MARK

Divisions of Biometry and Cardiology, Box 3363, Duke University Medical Center, Durham, North Carolina 27710, U.S.A.

SUMMARY

Multivariable regression models are powerful tools that are used frequently in studies of clinical outcomes. These models can use a mixture of categorical and continuous variables and can handle partially observed (censored) responses. However, uncritical application of modelling techniques can result in models that poorly fit the dataset at hand, or, even more likely, inaccurately predict outcomes on new subjects. One must know how to measure qualities of a model's fit in order to avoid poorly fitted or overfitted models. Measurement of predictive accuracy can be difficult for survival time data in the presence of censoring. We discuss an easily interpretable index of predictive discrimination as well as methods for assessing calibration of predicted survival probabilities. Both types of predictive accuracy should be unbiasedly validated using bootstrapping or cross-validation, before using predictions in a new data series. We discuss some of the hazards of poorly fitted and overfitted regression models and present one modelling strategy that avoids many of the problems discussed. The methods described are applicable to all regression models, but are particularly needed for binary, ordinal, and time-to-event outcomes. Methods are illustrated with a survival analysis in prostate cancer using Cox regression.

1. INTRODUCTION

Accurate estimation of patient prognosis is important for many reasons. First, prognostic estimates can be used to inform the patient about likely outcomes of her disease. Second, the physician can use estimates of prognosis as a guide for ordering additional tests and selecting appropriate therapies. Third, prognostic assessments are useful in the evaluation of technologies; prognostic estimates derived both with and without using the results of a given test can be compared to measure the incremental prognostic information provided by that test over what is provided by prior information.[1] Fourth, a researcher may want to estimate the effect of a single factor (for example, treatment given) on prognosis in an observational study in which many uncontrolled confounding factors are also measured. Here the simultaneous effects of the uncontrolled variables must be controlled (held constant mathematically if using a regression model) so that the effect of the factor of interest can be more purely estimated. An analysis of how variables (especially continuous ones) affect the patient outcomes of interest is necessary to

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
 ISBN: 0-470-02365-1

ascertain how to control their effects. Fifth, prognostic estimation is useful in designing randomized clinical trials. Both the decision concerning which patients to randomize and the design of the randomization process (for example, stratified randomization using prognostic factors) are aided by the availability of accurate prognostic estimates before randomization.[2] Lastly, accurate prognostic models can be used to test for differential therapeutic benefit or to estimate the clinical benefit for an individual patient in a clinical trial, taking into account the fact that low-risk patients must have less absolute benefit (lower change in survival probability).[3]

To accomplish these objectives, analysts must create prognostic models that accurately reflect the patterns existing in the underlying data and that are valid when applied to comparable data in other settings or institutions. Models may be inaccurate due to violation of assumptions, omission of important predictors, high frequency of missing data and/or improper imputation methods, and especially with small datasets, overfitting. The purpose of this paper is to review methods for examining lack of fit and detection of overfitting of models and to suggest guidelines for maximizing model accuracy. Section 2 covers initial steps such as imputation of missing data, pre-specification of interactions, and choosing the outcome model. Section 3 has an overview of the need for data reduction. In Section 4, we discuss the process of checking whether a hypothesized model fits the data. In Section 5, measures of predictive accuracy are covered. These are not directly related to lack of fit but rather to the ability of the model to discriminate and be well calibrated when applied prospectively. Section 6 covers model validation and demonstrates advantages of resampling techniques. Section 7 provides one modelling strategy that takes into account ideas from earlier sections and lists some miscellaneous concerns. Most of the methods presented here can be used with any regression model. Section 8 briefly describes some statistical software useful in carrying out the strategy summarized in Section 7. Section 9 has a detailed case study using a Cox regression model for time until death in a clinical trial studying prostate cancer.

2. PRELIMINARY STEPS

Before analyses begin, the researcher must specify the relationships of interest and define and assemble the response variable and the potential predictors. At this point a frequent problem is the extent of missing data. Some methods of dealing with missing data are given in References 4–7. Deletion of cases with missing predictors causes bias and increased variance. Even though caution should be taken when imputing missing values, it is usually better to estimate selected data values than to delete an entire subject's record. Simple methods of imputation include the use of the median, mean, or mode for missing values. This method is biased and inefficient when predictors are correlated with one another.[4] Deriving customized regression models for predicting each predictor from all other predictors is a better method. Kuhfeld[8] has implemented a general imputation method that allows predictors to be non-linearly (and even non–monotonically) related to one another. This method has been modified by Harrell and implemented in the S-Plus `transcan` function (Section 8), which yields stable imputations even when the fraction of missing values is quite large. In some cases, surrogate predictors, not intended to enter the model directly, are assembled to assist in imputing missing predictors in the model.

It is important that maximum information be extracted from predictors and response. Because of this and because of problems with data reliability, when one has a choice of describing a concept with a categorical variable or a continuous one, the continuous one is preferred. Subject matter knowledge should guide the selection of candidate predictors. Early deletion of those with little chance of being predictive or of being measured reliably will result in models with less overfitting and greater generalizability.

Plausible interactions should be carefully chosen because of problems of multiple parameters (see reference 9 for additional thoughts on interactions). Certain types of interactions that have frequently been found to be important in predicting clinical outcomes and thus may be pre-specified are:

1. Interactions between treatment and the severity of disease being treated. Patients with little disease have little opportunity to receive benefit.
2. Interactions involving age and risk factors. Old subjects are generally less affected by risk factors. They have been robust enough to survive to their current age with risk factors present.
3. Interactions involving age and type of disease. Some diseases are incurable and have the same prognosis regardless of age. Others are treatable or have less effect on younger patients.
4. Interactions between a measurement and the state of a subject during a measurement. For example, left ventricular function measured at rest may have less predictive value and thus have a smaller slope versus outcome than function measured during stress.
5. Interactions between calendar time and treatment. Some treatments evolve or their effectiveness improves with staff training.
6. Interactions between quality and quantity of a symptom.

Careful fitting of a statistical model is essential so that interactions, if present, represent biologic phenomena rather than general lack of fit of the model.

A tentative choice of the statistical model is sometimes based on previous distributional examinations, but it is frequently based on maximizing how available information is used. Binary and ordinal logistic models[10–13] are frequently used for discrete completely assessed outcomes, and the Cox proportional hazards model[14,15] and parametric survival models[16] are frequently used for censored time-to-event data. It is quite common to change the model after initial modelling of predictors, because only then can adjusted distributional properties of Y and joint properties of X and Y be assessed (Section 4.3).

3. DATA REDUCTION

Multivariable statistical models when developed carefully are excellent tools for making prognostic predictions. However, when the assumptions of a model are grossly violated or when a model is used unwisely for a given patient sample, the performance of the model may be poor. For example, when the analyst has fitted not only real trends that further data would support, but in addition has fitted idiosyncrasies in the particular dataset by analysing too many variables, the model may predict inaccurately for a new group of patients. Only with appropriate model validation can an apparently accurate model be shown to be inaccurate.

In developing a set of predictions based on 100 patients, no analyst would divide the patients into 50 subgroups and quote the average outcome for each subgroup. Yet many articles have appeared in the clinical literature where 20–50 variables were analysed on 100 patients. Researchers apparently do not realize that when many predictor variables are analysed, variable screening based on statistical significance and stepwise variable selection involve multiple comparisons problems that lead to unreliable models. These methods are therefore not viable for data reduction (see Reference 17 for a condemnation of stepwise variable selection).

The situation is actually worse than merely considering the number of predictors. If the analyst used associations with Y to entertain non-linearities in the predictors or interaction terms, these constructed variables need to be counted (see Table II for an example). We speak of the total

predictor degrees of freedom (d.f.), p, as the total number of parameters (columns of the design matrix) examined during the course of analysis, excluding intercept term(s). If graphical or other informal analyses are used to guide the analysis, it is difficult to define p – one needs to estimate the effective number of parameters considered according to the flexibility of fits that were considered.[18] The quantity p is the effective number of parameters allowed for consideration, that is, the number of regression coefficients estimated formally or informally without algebraic restrictions.

To enhance the accuracy of a model, the number of variables used must be reduced or the model must be simplified unless the sample is large. Unless a formal penalized estimation technique is used,[19] multiple comparisons problems that arise from 'peeking' at the outcome variable must be eliminated; data reduction methods must be used that do not utilize the outcome variable. Harrell *et al.*[20] discussed some available data reduction methods and two regression modelling strategies based on these methods that yield reliable models. They suggest as a rough rule of thumb that in order to have predictive discrimination that validates on a new sample, no more than $m/10$ predictor d.f. p should be examined to fit a multiple regression model, where m is the number of uncensored event times (for example, deaths) in the training sample (the sample used in fitting the model). For binary outcomes m is the number of patients in the less frequent outcome category. If $p > m/10$, a data reduction technique such as principal components, variable clustering, or deriving clinical summary indexes[20–23] should be used until the number of summary variables to use as candidates in the regression analysis is less than $m/10$.

Smith *et al.*[24] found in one series of simulations that the expected error* in Cox model predicted 5-year survival probabilities was below 0·05 when $p < m/20$ for 'average' subjects and below 0·10 when $p < m/20$ for 'sick' subjects. For 'average' subjects, $m/10$ was adequate for preventing expected errors $> 0{\cdot}1$.

Better and more general than any of these rules is the reduction of d.f. using a shrinkage method (Section 5.4).

4. VERIFYING MODEL ASSUMPTIONS: CHECKING LACK OF FIT

4.1. Linearity assumption

In their simplest forms, all usual regression models assume that for a certain scale of Y, each predictor variable X is linearly related to Y. In the logistic regression model for binary responses, the initial assumption is that an X is linearly related to the log odds of response ($\log[P/(1-P)]$, where P is the probability of response) for patients subgrouped by values of X. In the Cox proportional hazards survival model, one initially assumes that at each time t, $\log[-\log(S(t))]$ and equivalently $\log\lambda(t)$ are linearly related to X, where $S(t)$ is the probability of surviving until time t and $\lambda(t)$ is the hazard function or instantaneous event rate at time t. It is easy to envision cases where strong violations in the linearity assumption (say a U-shaped age relationship) will result in erroneous predictions.

A direct way to check the linearity assumption, and to determine how to transform a specific X if necessary, involves expanding X into multiple terms that can flexibly fit any smooth relationship. The extra terms can be statistically tested to assess the adequacy of a linear relationship, and the terms *in toto* can estimate the true transformation of X that would result in

* Absolute difference between predicted and actual 5-year survival probabilities in a simulation study with known survival functions

a linear relationship with Y. A common choice of expansion is to add X^2 and perhaps higher powers of X to the model. A more flexible approach is the use of piecewise linear regression or piecewise cubic polynomials (spline functions). See references 25–27 for methods of fitting such functions.

As an alternative, smoothed residual plots can be used to determine the functional form for each predictor. For binary logistic models, smoothed partial residual plots[13,28,29] are useful, and for the Cox model, smoothed martingale residuals plots detect regression shape departures.[30] Partial residuals in logistic models are particularly computationally efficient, as the analyst can fit a simple model that is linear in all predictors and then use the residuals to obtain estimates of the *true* functional forms. However, the plot for each predictor does assume that the other predictors operate linearly and that all predictors are additive (see below). The usual martingale residual plot for the Cox model provides an estimate of the *departure* from linearity for the predictor.

4.2. Additivity assumption

A further assumption of most regression models is additivity of effects of the predictors (lack of interaction). Interactions can be tested and described by adding cross-product terms. It must be borne in mind that interactions can take the form of a change in shape (for example, linear age relationship for males, quadratic for females), so the cross-products needed in the model are not always simple ones.

The number of possible cross-product terms is usually so large (especially when variables have non-linear or multiple dummy variable components) that the predictors to check for additivity must usually be specified before examining the data. Otherwise, type I errors and overfitting will be significant problems. A compromise solution is to do pooled interaction tests. For example, in a model with predictors age, sex, and dose, one may test all second-order interactions involving age, all interactions involving sex, and all involving dose. A combined test of all two-way interactions is also useful. If a pooled test is not significant, it may be unwise to pursue significant component interactions.

4.3. Distributional assumption

The previous sections dealt with the proper specification of the X-structure of the model. Once the analyst has determined which predictors are to be used and how they should be represented in the model, most models have distributional assumptions that also need verification. The Cox model does not assume anything about the survival function $S(t)$ across t for an individual, but it does assume how survival curves for different subjects are related. Specifically, it assumes that $\log[-\log(S(t))]$ for different subjects are equidistant over time, or equivalently that hazard functions for any two subjects are proportional over time. This proportional hazards assumption can be checked using smoothed plots of a special type of residual from the model called the Schoenfeld residual.[31,32] It can also be checked using hazard ratio plots, plots of modelled versus stratified estimates,† and several other methods.[33] Unlike the Cox model, fully parametric models (for example, Weibull or log-normal survival models) have a distributional assumption even when there are no covariables. If the form of $S(t)$ does not fit the data for these models, estimates of $S(t)$ will be inaccurate.

† That is, a Cox model is fitted with the variable in question appearing as a covariate for which regression coefficient(s) are estimated, then a second model is fitted where that variable is used as a stratification factor that modifies the underlying survival function (but which does not have regression coefficients).

5. QUANTIFYING PREDICTIVE ACCURACY

There are at least three uses of measures of predictive accuracy:

1. To quantify the utility of a predictor or model to be used for prediction or for screening to identify subjects at increased risk of a disease or clinical outcome.‡
2. To check a given model for overfitting (fitting noise resulting in unstable regression coefficients) or lack of fit (improper model specification, omitted predictors, or underfitting). More will be said about this later.
3. To rank competing methods or competing models.

The measures discussed below may be applied to the assessment of a predictive model using the same sample on which the model was developed. However, this assessment is seldom of interest, as only the most serious lack of fit will make a model appear not to fit on the sample for which it was tailor-made. Of much greater value is the assessment of accuracy on a separate sample or a bias-corrected estimate of accuracy on the training sample. This assessment can detect gross lack of fit as well as overfitting, whereas the *apparent* accuracy from the original model development sample does not allow one to quantify overfitting. Section 6 discusses how the indexes described below may be estimated fairly using a validation technique.

5.1. General notions

In the simplest case, when the response being predicted is a continuous variable that is measured completely (as distinct from *censored* measurements caused by termination of follow-up before all subjects have had the outcome of interest), one commonly used measure of predictive accuracy is the *expected squared error* of the estimate. This quantity is defined as the expected squared difference between predicted and observed values, that is, the average squared difference between predicted and observed values if the experiment were repeated infinitely often and new estimates were made at each replication. The expected squared error can also be expressed as the square of the *bias* of the estimate plus the *variance* of the estimate. Here bias refers to the expected value of the estimate minus the quantity being estimated, such as the mean blood pressure. The expected squared error is estimated in practice by the usual mean squared error.

There are two other terms for describing the components of predictive accuracy: *calibration* and *discrimination*. Calibration refers to the extent of bias. For example, if the average predicted mortality for a group of similar patients is 0·3 and the actual proportion dying is 0·3, the predictions are well calibrated. Discrimination measures a predictor's ability to separate patients with different responses. A weather forecaster who predicts a 0·15 chance of rain every day of the year may be well calibrated in a certain locality if the average number of days with rain is 55 per year, but the forecasts are uninformative. A discriminating forecaster would be one who assigns a wide distribution of predictions and whose predicted risks for days where rain actually occurred are larger than for dry days. If a predictive model has poor discrimination, no adjustment or

‡ Often one wishes to designate a model as 'minimally acceptable' on the basis of some statistic, but in many cases it is only possible to judge a model's accuracy relative to another model. For example, a model for the probability of death after open heart surgery may yield predicted probabilities that range from 0·001 to 0·1, so the model will not have a high correlation (say 0·13) between predicted probability and observed outcome, but it may still be useful. If that model does not fully adjust for patient risk factors, it may be inadequate for adjusting for case mix when comparing mortalities among several hospitals. A more sensitive model with a correlation of, say, 0·135 may adjust away apparent differences in mortality among hospitals.

calibration can correct the model. However, if discrimination is good, the predictor can be calibrated without sacrificing the discrimination (see Section 6 for a method for calibrating predictions without needing more data). Here, calibrating predictions means modifying them, without changing their rank order, such that the predictions are perfectly calibrated. van Houwelingen and le Cessie[34] present extensive information on predictive accuracy and model validation.

5.2. Continuous uncensored outcomes

Discrimination is related to the expected squared error and to the correlation between predicted and observed responses. In the case of ordinary multiple linear regression, discrimination can be measured by the squared multiple correlation coefficient R^2, which is defined by

$$R^2 = 1 - (n - p)\text{MSE}/(n - 1)S_Y^2, \tag{1}$$

where n is the number of patients, p is the number of parameters estimated, MSE is the mean squared error of prediction $(\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2/(n - p)$, $\hat{Y}$ = predicted Y), and S_Y^2 is the sample variance of the dependent variable. When $R^2 = 1$, the model is perfectly able to separate all patient responses based on the predictor variables, and MSE = 0.

For a continuous uncensored response Y, calibration can be assessed by a scatter plot of $\hat{Y}$ (predicted Y) versus Y, optionally using a non-parametric smoother to make trends more evident.

5.3. Discrete or censored outcomes

When the outcome variable is dichotomous and predictions are stated as probabilities that an event will occur, calibration and discrimination are more informative than expected squared error alone in measuring accuracy.

One way to assess calibration of probability predictions is to form subgroups of patients and check for bias by comparing predicted and observed responses (reference 29, pp. 140–145). For example, one may group by deciles of predicted probabilities and plot the mean response (proportion with the outcome) versus the mean prediction in the decile group. However, the groupings can be quite arbitrary. Another approach is to use a smoother such as the 'super smoother'[35] or a scatterplot smoother[36] to obtain a non-parametric estimate of the relationship between $\hat{Y}$ and Y. Such smoothers work well even when Y is binary. The resulting smoothed function is a nonparametric calibration or reliability curve. Smoothers operate on the raw data $(\hat{Y}, Y)$ and do not require grouping $\hat{Y}$, but they do require one to choose a smoothing parameter or bandwidth.

As an example, consider a 7-variable binary logistic regression model to predict the probability that a certain disease is present. The model was developed on a simulated 200-subject dataset of whom 93 had a final diagnosis that is positive. While fixing the intercept and 7 regression coefficients estimated from the training sample, predictive probabilities of disease were computed for each of 200 subjects in a separate sample, of whom 104 had the disease. The non-parametric calibration curve was estimated using a local least squares scatterplot smoother[36] with the S-Plus function **lowess**,[37] using the 'no iteration' option. The smoothed calibration graph is shown in Figure 1. Also shown is the proportion of patients with disease, grouped by intervals of predicted probability each containing 50 patients.

Note the typical regression to the mean effect caused by overfitting: predicted probabilities in the range of 0·3 to 0·5 are too low. Actual probabilities are closer to the mean (104/200 = 0·52).

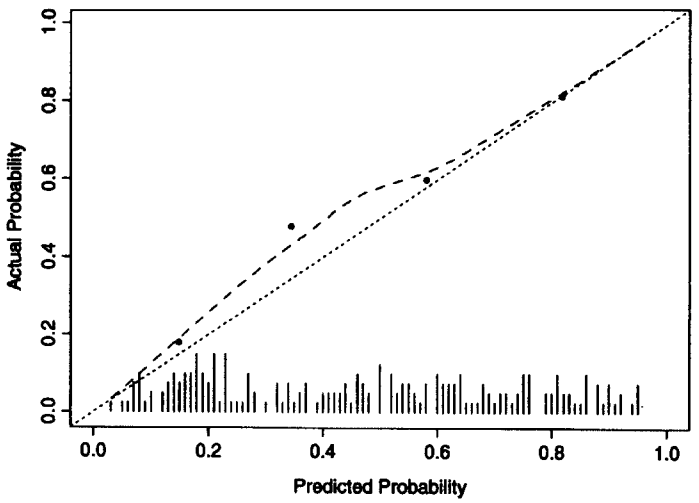


Figure 1. Smooth non-parametric calibration curve (dashed line), subgroup estimates (dots), and ideal relationship (dotted line). The distribution of predicted probabilities is shown above the x-axis. 'Actual probability' is an unbiased estimate of the true probability of response given the level of the predicted probability

When Y is binary and $\hat{Y}$ is the predicted probability that $Y = 1$ versus $Y = 0$, the Brier score[38] or average $(Y - \hat{Y})^2$ is a commonly used mean squared error-type measure of predictive accuracy.

For survival models, one may choose one or more times $(t_1, t_2, \ldots, t_k)$, and plot the predicted probability of surviving until each t_j versus the actual fraction of patients surviving past t_j. The problem here is that we cannot define $Y_i = 1$ if patient i survives past time t_j and then plot the mean Y (by deciles of $\hat{Y}$ or using a smoother) against the mean $\hat{Y}$, since subjects not followed until time t_j are censored, that is, their final outcome status is unknown. One solution is to divide the sample into intervals of $\hat{Y}$ so that there are 50 subjects in each interval of predicted survival, and then plot the mean $\hat{Y}$ within each interval versus the Kaplan–Meier[39] survival estimate at time t_j.

5.4. Shrinkage

Shrinkage is the flattening of the plot of (predicted, observed) away from the 45° line, caused by overfitting. It is a concept related to regression to the mean. One can estimate the amount of shrinkage present (using external validation) or the amount likely to be present (using bootstrapping, cross-validation or simple heuristics). A shrinkage coefficient can be used to quantify overfitting or one can go a step further and use the coefficient to re-calibrate the model. Shrinkage can be defined as a multiplier γ of $X\hat{\beta}$ (excluding intercept(s)) needed to make $\gamma X\hat{\beta}$ perfectly calibrated for future data. The heuristic shrinkage estimator of van Houwelingen and le Cessie[34] (see also reference 40) is

$$\hat{\gamma} = \frac{\text{model } \chi^2 - p}{\text{model } \chi^2}, \tag{2}$$

where p is the number of regression parameters (here excluding any intercept(s) but including all non-linear and interaction effects) and the model χ^2 is the total likelihood ratio χ^2 statistic (computed using the full set of p parameters) for testing whether any predictors are associated

with Y.[§] For linear regression, van Houwelingen and le Cessie's heuristic shrinkage estimate reduces to the ratio of the adjusted R^2 to the ordinary R^2 (derivable from reference 34, Eq. 70).

As an example, suppose that an analyst has considered 10 predictor variables, 6 of which were allowed to enter the model non-linearly (with 2 non-linear terms for each), and tested 8 interaction terms, for a total of 30 degrees of freedom. The model χ^2 is 100 for the full model fit with $p = 30$ d.f. The expected shrinkage is 0·70, indicating that about 0·3 of the model fit is 'noise'. The 'final model' obtained from forward variable selection contains only 3 significant coefficients and has $\chi^2 = 81$, but overfitting is quantified using the 30 candidate d.f. In this example, the number of variables, transformations, and interactions tried was too many for the sample size, and the resulting model is expected to be unstable. As a rough estimate, 0·3 of what was learned from developing the model was really non-replicable noise.

For mild overfitting in the case where the model is needed only to rank likely outcomes and not predict absolute risks, shrinking the regression coefficients will not help since it will not increase real discrimination. If the model is badly overfitted, the model may actually have negative (worse than random) discrimination on new data, and it will have poor calibration. The following heuristic strategy can then be used to determine whether data reduction is likely to result in a model that has any discrimination and how much reduction is required to yield reliable non-shrunken predictions.

First, fit a full model with all candidate variables, non-linear terms, and hypothesized interactions. Let p denote the number of parameters in this model, aside from any intercept(s). Let LR denote the likelihood ratio χ^2 for this full model. The estimated shrinkage is $(\text{LR} - p)/\text{LR}$. If this falls below 0·85, for example, we may be concerned. Let q denote the regression degrees of freedom for a reduced model. In a 'best case', the variables removed to arrive at the reduced model would have no association with Y. The expected value of the χ^2 statistic for testing those variables would then be $p - q$. The shrinkage for the reduced model is then on average $[\text{LR} - (p - q) - q]/[\text{LR} - (p - q)]$. Solving for q gives $q \leqslant (\text{LR} - p)/9$. Therefore, reduction of dimensionality down to q degrees of freedom would be expected to achieve < 10 per cent shrinkage. With these assumptions, there is no hope that a reduced model would have acceptable calibration unless $\text{LR} > p + 9$. If the information explained by the omitted variables is less than one would expect by chance (for example, their total χ^2 is extremely small), a reduced model could still be beneficial, as long as the conservative bound $(\text{LR} - q)/\text{LR} \geqslant 0{\cdot}9$ or $q \leqslant \text{LR}/10$ were achieved. This conservative bound assumes that no χ^2 is lost by the reduction, that is, that the final model $\chi^2 \approx$ LR. This is unlikely in practice, since the data reduction *must* be only X-driven.

As an example, suppose that a binary logistic model is being developed from a sample containing 45 events on 150 subjects. A 10:1 events: d.f. rule suggests we can analyse 4·5 degrees of freedom. The analyst wishes to analyse age, sex, and 10 other variables. It is not known whether interaction between age and sex exists, and whether age is linear. A restricted cubic spline is fitted with 4 knots (requiring two non-linear terms), and a linear interaction is allowed between age and sex. These two variables then need $3 + 1 + 1 = 5$ degrees of freedom. The other 10 variables are assumed to be linear and to not interact with themselves or age and sex. There is a total of 15 d.f. The full model with 15 d.f. has LR = 50. Expected shrinkage from this model is

[§] When stepwise fitting is done, the definition of p is confusing. Many analysts act as if the final model chosen with stepwise variable selection was pre-specified, whether interpreting R^2, confidence limits, or P-values. For estimating the likely shrinkage, it has been shown that p is much closer to the number of candidate d.f. than to the number of parameters fitted in a 'final' model.[40] On a similar note, reference 18 showed how to adjust a linear test of association for having done a test of quadratic effect, concluding that testing the single d.f. statistic for association as if it had 2 d.f. is nearly optimal.

$(50 - 15)/50 = 0{\cdot}7$. Since LR $> 15 + 9 = 24$, some reduction *might* yield a better validating model. Reduction to $q = (50 - 15)/9 \approx 4$ d.f. would be necessary, assuming the reduced LR is about $50 - (15 - 4) = 39$. In this case the 10:1 rule yields about the same value for q. The analyst may be forced to assume that age is linear, modelling 3 d.f. for age and sex. The other 10 variables would have to be reduced to a single variable using principal components or another scaling technique. This single variable may not be interpretable, but using a single score is better than deleting all 10 variables from consideration. If the goal of the analysis is to make a series of hypothesis tests (adjusting P-values for multiple comparisons) instead of to predict future responses, the full model would have to be used.

Bootstrapping[34] and cross-validation[41] may also be used to estimate shrinkage factors. As mentioned above, shrinkage estimates are useful in their own right for quantifying overfitting, and they are also useful for 'tilting' the predictions so that the (predicted, observed) plot does follow the 45° line, by multiplying all of the regression coefficients by $\hat{\gamma}$. However, for the latter use it is better to follow a more rigorous approach such as penalized maximum likelihood estimation,[19] which allows the analyst to shrink different parts (for example, non-linear terms or interactions) of the equation more than other parts.[42]

5.5. General discrimination index

Discrimination can be defined more uniquely than calibration. It can be quantified with a measure of correlation without requiring the formation of subgroups or requiring smoothing.

When dealing with binary dependent variables or continuous dependent variables that may be censored when some patients have not suffered the event of interest, the usual mean squared error-type measures do not apply. A c (for *concordance*) index[1] is a widely applicable measure of predictive discrimination – one that applies to ordinary continuous outcomes, dichotomous diagnostic outcomes, ordinal outcomes, and censored time until event response variables. This index of predictive discrimination is related to a rank correlation between predicted and observed outcomes. It is a modification of the Kendall–Goodman–Kruskal–Somers type rank correlation index[43] and was motivated by a modification of Kendall's τ by Brown *et al.*[44] and Schemper.[45]

The c index is defined as the proportion of all usable patient pairs in which the predictions and outcomes are concordant. The c index measures predictive information derived from a set of predictor variables in a model. In predicting the time until death, c is calculated by considering all possible pairs of patients, at least one of whom has died. If the predicted survival time is larger for the patient who lived longer, the predictions for that pair are said to be concordant with the outcomes. If one patient died and the other is known to have survived at least to the survival time of the first, the second patient is assumed to outlive the first. When predicted survivals are identical for a patient pair, $\frac{1}{2}$ rather than 1 is added to the count of concordant pairs in the numerator of c. In this case, one is still added to the denominator of c (such patient pairs are still considered usable). A patient pair is unusable if both patients died at the same time, or if one died and the other is still alive but has not been followed long enough to determine whether she will outlive the one who died.

Instead of using the predicted survival time to calculate c, the predicted probability of surviving until any fixed time point can be used equivalently, as long as the two estimates are one-to-one functions of each other. This holds for example if the proportional hazards assumption is satisfied.

For predicting binary outcomes such as the presence of disease, c reduces to the proportion of all pairs of patients, one with and one without the disease, in which the patient having the disease had the higher predicted probability of disease. As before, pairs of patients having the same

predicted probability get $\frac{1}{2}$ added to the numerator. The denominator is the number of patients with disease multiplied by the number without disease. In this binary outcome case, c is essentially the Wilcoxon–Mann–Whitney statistic for comparing predictions in the two outcome groups, and it is identical to the area under a receiver operating characteristic (ROC) curve.[46,47] Liu and Dyer[48] advocate the use of rank association measures such as c in quantifying the impact of risk factors in epidemiologic studies.

The c index estimates the probability of concordance between predicted and observed responses. A value of 0·5 indicates no predictive discrimination and a value of 1·0 indicates perfect separation of patients with different outcomes. For those who prefer instead a rank correlation coefficient ranging from -1 to $+1$ with 0 indicating no correlation, Somers' D rank correlation index is derived by calculating $2(c - 0{\cdot}5)$. Either c or the rank correlation index can be used to quantify the predictive discrimination of any quantitative predictive method, whether the response is continuous, ordinal, or binary.

Even though rank indexes such as c are widely applicable and easily interpretable, they are not sensitive for detecting small differences in discrimination ability between two models. This is due to the fact that a rank method considers the (prediction, outcome) pairs (0·01, 0), (0·9, 1) as no more concordant than the pairs (0·05, 0), (0·8, 1). A more sensitive likelihood-ratio χ^2-based statistic that reduces to R^2 in the linear regression case may be substituted.[49–51] Korn and Simon[52] have a very nice discussion of various indexes of accuracy for survival models.

6. MODEL VALIDATION METHODS

As mentioned before, examination of the *apparent* accuracy of a multivariable model using the training dataset is not very useful. The most stringent test of a model (and of the entire data collection system) is an external validation – the application of the 'frozen' model to a new population. It is often the case that the failure of a model to validate externally could have been predicted from an honest (unbiased) 'internal' validation. In other words, it is likely that many clinical models which failed to validate would have been found to fail on another series of subjects from the original source, because overfitting is such a common problem. The principal methods for obtaining nearly unbiased internal assessments of accuracy are *data-splitting*,[53] *cross-validation*[54] and *bootstrapping*.[54–58] In data-splitting, a random portion, for example $\frac{2}{3}$, of the sample is used for all model development (data transformations, stepwise variable selection, testing interactions, estimating regression coefficients, etc.). That model is 'frozen' and applied to the remaining sample for computing calibration statistics, c, etc. The size of the validation sample must be such that the relationship between predicted and observed outcomes can be estimated with good accuracy, and the remaining data are used as the training (model development) sample. Data-splitting is simple, because all the modelling steps, which may include subjective judgements, are only done once. Data-splitting also has an advantage when it is feasible to make the single split with respect to geographical location or time, resulting in a more stringent validation that demonstrates generalizability. However, in addition to severe difficulties listed below, data splitting does not validate the final model, if one desires to recombine the training and test data to derive a model for others to use.

Cross-validation is repeated data-splitting. To obtain accurate estimates using cross-validation, more than 200 models may need to be developed and tested,[54] with results averaged over the 200 repetitions. For example, in a sample of size $n = 1000$, the modelling process (all components of it!) could be done 400 times, leaving out a random 50 subjects each time and developing the model on the 950 remaining subjects. The benefits of cross-validation over data-splitting are

clear; the size of the training samples can be much larger, so less data are discarded from the estimation process. Secondly, cross-validation reduces variability by not relying on a single sample split.

Efron has shown that cross-validation is relatively inefficient due to high variation of accuracy estimates when the entire validation process is repeated.[54] Data-splitting is far worse; the indexes of accuracy will vary greatly with different splits. Bootstrapping is an alternative method of internal validation that involves taking a large number of samples with replacement from the original sample. Bootstrapping provides nearly unbiased estimates of predictive accuracy that are of relatively low variance, and fewer model fits are required than cross-validation. Bootstrapping has an additional advantage that the entire dataset is used for model development. As others have shown, data are too precious to waste.[59,60]

Suppose that we wish to estimate the expected value (for new patient samples similar to the derivation sample) of the Somers' D rank correlation coefficient between predicted and observed survival time. The following steps can be used (see references 55, 58 and 60 for the basic method when applied to binary outcomes):

1. Develop the model using all n subjects and whatever stepwise testing is deemed necessary. Let D_{app} denote the *apparent D* from this model, i.e., the rank correlation computed on the same sample used to derive the fit.
2. Generate a sample of size n with replacement from the original sample (for both predictors and the response).
3. Fit the full or possibly stepwise model, using the same stopping rule as was used to derive D_{app}.
4. Compute the apparent D for this model on the bootstrap sample with replacement. Call it D_{boot}.
5. 'Freeze' this reduced model, and evaluate its performance on the original dataset. Let D_{orig} denote the D.
6. The optimism in the fit from the bootstrap sample is $D_{boot} - D_{orig}$.
7. Repeat steps 2 to 6 100–200 times.
8. Average the optimism estimates to arrive at O.
9. The bootstrap corrected performance of the original stepwise model is $D_{app} - O$. This difference is a nearly unbiased estimate of the *expected value* of the external predictive discrimination of the process which generated D_{app}. In other words, $D_{app} - O$ is an *honest* estimate of *internal* validity, penalizing for overfitting.

As an example, suppose we want to validate a stepwise Cox model developed from, say, a sample of size $n = 300$ with 30 events. The candidate regressors are age, age^2, sex, mean arterial blood pressure (MBP), and a non-linear interaction between age and sex with the terms age × sex and age^2 × sex. MBP is assumed to be linear and additive. Denote these variables by the numbers 1–6. The model χ^2 is 45 with 6 d.f., so the approximate expected shrinkage is $\frac{45-6}{45} = 0{\cdot}87$, or 0·13 overfitting, so some caution needs to be exercised in using the estimated model coefficients and hence in using extreme predicted survival probabilities without calibration (shrinkage). The D for the full model is 0·42. A step-down variable selection using Akaike's information criterion (AIC)[34,61] as a stopping rule (χ^2 for set of variables tested $> 2 \times$ d.f.) resulted in a model with the variables age, age^2, sex, age × sex. The reduced model had $D = 0{\cdot}39$, a typical loss due to deleting marginally important but statistically insignificant variables. Two-hundred bootstrap repetitions are done, repeating the variable selection for each sample using the same stopping rule. We want to detect whether the $D = 0{\cdot}39$ is likely to validate in a new series of subjects from the same population. The first five samples might yield the results shown in Table I.

Table I. Example validation of predictive discrimination

Re-sample	D_{boot} Full model	Variables retained	D_{boot} Reduced model	D_{orig}	Optimism
1	0·45	1, 2, 3, 5, 6	0·44	0·37	0·07
2	0·46	1, 2	0·34	0·30	0·04
3	0·42	1, 2, 3, 4	0·37	0·34	0·03
4	0·43	1, 2, 3, 5	0·42	0·39	0·03
5	0·41	1, 3, 4	0·39	0·37	0·02

The average optimism is 0·038, so the bootstrap estimate of the expected validation of D_{app} is 0·39 − 0·038 = 0·352. The analyst may or may not be worried about the 0·038 overfitting, but the best estimate of predictive discrimination is $D = 0·352$ – this is a better estimate of the likely 'external' validation accuracy than is 0·39 if all other aspects of the study design remain constant. The $D = 0·352$ is the honest estimate of predictive accuracy that should be quoted when the researchers document the accuracy of the reduced model that was developed on the entire dataset using a stepwise variable selection algorithm.

It is usually informative to repeat the bootstrap validation with and without stepwise variable selection. Usually, the amount of predictive information lost by deleting marginal variables is not offset by the decreased optimism of the stepwise model. One way to demonstrate this point is to observe how often 'insignificant' clinical predictors have clinically sensible signs on their regression coefficients. Stepwise variable selection, which requires binary decisions about the inclusion of variables (unlike shrinkage), causes information to be lost.[2]

The same strategy can be used to estimate the over-optimism in an R^2 measure[49] from the original model fit. For estimating the prediction error at time t in a survival model, similar steps could also be used. Instead of validating a correlation D, we substitute for example the statistic D = difference between mean predicted 2-year survival probability and Kaplan–Meier 2–year survival estimate. The survival estimates are made by, say, deciles of predicted 2-year survival from the original model fit using the following steps, for example:

1. Develop the model using all subjects.
2. Compute cut points on predicted survival at 2 years so that there are m patients within each interval ($m = 50$ or 100 typically).
3. For each interval of predicted probability, compute the mean predicted 2-year survival and the Kaplan–Meier 2-year survival estimate for the group.
4. Save the apparent errors – the differences between mean predicted and Kaplan–Meier survival.
5. Generate a sample with replacement from the original sample.
6. Fit the full model.
7. Do variable selection and fit the reduced model.
8. Predict 2-year survival probability for each subject in the bootstrap sample.
9. Stratify predictions into intervals using the previously chosen cut points.
10. Compute Kaplan–Meier survival at 2 years for each interval.
11. Compute the difference between the mean predicted survival within each interval and the Kaplan–Meier estimate for the interval.
12. Predict 2-year survival probability for each subject in the original sample using the model developed on the sample with replacement.

13. For the same cut points used before, compute the difference in the mean predicted 2-year survival and the corresponding Kaplan–Meier estimates for each group in the original sample.
14. Compute the differences in the differences between the bootstrap sample and the original sample.
15. Repeat steps 5 to 14 100–200 times.
16. Average the 'double differences' computed in step 14 over the 100–200 bootstrap samples. These are the estimates of over-optimism in the apparent error estimates.
17. Add these over-optimism estimates to the apparent errors in the original sample to obtain bias-corrected estimates of predicted versus observed, that is, to obtain a bias- or overfitting-corrected calibration curve.

7. SUMMARY OF MODELLING STRATEGY

1. Assemble accurate, pertinent data and as large a sample as possible. For survival time data, follow-up must be sufficient to capture enough events as well as the clinically meaningful phases if dealing with a chronic disease.
2. Formulate focused clinical hypotheses that lead to specification of relevant candidate predictors, the form of expected relationships, and possible interactions.
3. Discard observations having missing Y after characterizing whether they are missing at random.‖ See reference 62 for a study of imputation of Y when it is not missing at random.
4. If there are any missing Xs, analyse factors associated with missingness. If the fraction of observations that would be excluded due to missing values is very small, or one of the variables that is sometimes missing is of overriding importance, exclude observations with missing values¶. Otherwise impute missing Xs using individual predictive models that take into account the reasons for missing, to the extent possible.
5. If the number of terms fitted *or* tested in the modelling process (counting non-linear and cross-product terms) is too large in comparison with the number of outcomes in the sample, use data reduction (ignoring Y)[20–23] until the number of remaining free variables needing regression coefficients is tolerable. Assessment of likely shrinkage (overfitting) can be useful in deciding how much data reduction is adequate. Alternatively, build shrinkage into the initial model fitting.[19]
6. Use the entire sample in the model development as data are too precious to waste. If steps listed below are too difficult to repeat for each bootstrap or cross-validation sample, hold out test data from all model development steps which follow.
7. Check linearity assumptions and make transformations in Xs as needed.
8. Check additivity assumptions and add clinically motivated interaction terms.
9. Check to see if there are overly-influential observations.[30] Such observations may indicate overfitting, the need for truncating the range of highly skewed variables or making other pre-fitting transformations, or the presence of data errors.

‖ For survival time data, no observations should be missing on Y. They should only have curtailed follow-up.

¶ Alternatively, impute missing values for the predictor but perform secondary analyses later to estimate the strength of association between X and Y after deleting observations with that predictor imputed, as imputation will attenuate the relationship.

10. Check distributional assumptions and choose a different model if needed (in the case of Cox models, stratification or time-dependent covariables can be used if proportional hazards is violated).
11. Do limited backwards step-down variable selection.[63] Note that since stepwise techniques do not really address overfitting and they can result in a loss of information, full model fits (that is, leaving all hypothesized variables in the model regardless of *P*-values) are frequently more discriminating than fits after screening predictors for significance.[2,40] They also provide confidence intervals with the proper coverage, unlike models that are reduced using a stepwise procedure,[60,64,65] from which confidence intervals are falsely narrow. A compromise would be to test a *pre-specified* subset of predictors, deleting them if their total $\chi^2 < 2 \times$ d.f. If the χ^2 is that small, the subset would likely not improve model accuracy.
12. This is the 'final' model.
13. Validate this model for calibration and discrimination ability, preferably using bootstrapping. Steps 7 to 11 must be repeated for each bootstrap sample, at least approximately. For example, if age was transformed when building the final model, and the transformation was suggested by the data using a fit involving age and age^2, each bootstrap repetition should include both age variables with a possible step-down from the quadratic to the linear model based on automatic significance testing at each step.
14. If doing stepwise variable selection, present a summary table depicting the variability of the list of 'important factors' selected over the bootstrap samples or cross-validations. This is an excellent tool for understanding why data-driven variable selection is inherently ambiguous.
15. Estimate the likely shrinkage of predictions from the model, either using equation (2) or by bootstrapping an overall slope correction for the predictions.[34] Consider shrinking the predictions to make them calibrate better, unless shrinkage was built-in. That way, a predicted 0·4 mortality is more likely to validate in a new patient series, instead of finding that the actual mortality is only 0·2 because of regression to the mean mortality of 0·1.

8. SOFTWARE

Modern statistical software such as S-Plus[37] on UNIX workstations makes it quite feasible to perform the extensive calculations required to do the recommended model building steps. The first author has written a package of UNIX S-Plus functions called Design[66] that allow the analyst to perform all analyses mentioned here including tests of linearity, pooled interaction tests, model validation and graphical methods for interpreting models. Here are some examples:

```
# First find optimum transformations relating each predictor to each
# other, and use multiple regression in these transformations to
# impute missing values. Use shrinkage to avoid over-imputing
trans ← transcan( ~ age + cholesterol + sys.bp + weight, imputed = T, shrink = T)
cholesterol ← impute(trans, cholesterol) # impute missings
sys.bp ← impute(trans, sys.bp)
# Fit a Cox P.H. model allowing some interactions with age and
# nonlinearity in cholesterol and sys.bp using restricted cubic splines
# x = T, y = T means store data in fit for future bootstrapping
fit ← cph(Surv(fu.time, death) ~ age * (rcs(cholesterol) + rcs(sys.bp)) +
                    weight, x = T, y = T, surv = T, time.inc = 5)
anova(fit)                          # automatic pooled Wald tests
fastbw(fit)                         # fast backward step-down
```

Table II. Candidate predictors and d.f.

Predictor	Name	Number of parameters	Original levels
Dose of oestrogen	rx	3	placebo, 0·2, 1·0, 5·0 mg oestrogen
Age in years	age	3	
Weight index: wt(kg) − ht(cm) + 200	wt	3	
Performance rating	pf	2	normal, in bed <50% of time, in bed >50%, in bed always
History of cardiovascular disease	hx	1	present/absent
Systolic blood pressure/10	sbp	3	
Diastolic blood pressure/10	dbp	3	
Electrocardiogram code	ekg	5	normal, benign, rhythm disturbance, block, strain, old myocardial infarct, new MI
Serum haemoglobin (g/100 ml)	hg	3	
Tumour size (cm^2)	sz	3	
Stage/histologic grade combination	sg	3	
Serum prostatic acid phosphatase	ap	3	
Bone metastasis	bm	1	present/absent

```
# Next validate model, penalizing for backward stepdown variable selection
validate(fit, B = 100, bw = T)          # bootstrap validation of accuracy indexes
calibrate(fit, B = 100, bw = T, u = 5) # bias-corrected 5-yr survival calibration
plot(summary(fit))                      # plot hazard ratios with confidence limits
nomogram(fit)                           # draw nomogram displaying how model works
latex(fit)                              # typeset model equation
```

The Design library includes a function rcorr.cens for computing the general c–index, and the function val.prob which produced Figure 1 and also prints a variety of accuracy measures. For binary and ordinal logistic models and for ordinary linear models, Design has a general penalized maximum likelihood estimation facility. Design is available in the statlib repository (Internet address lib.stat.cmu.edu). transcan and impute are separate functions in statlib which work on UNIX as well as DOS Windows S-Plus. Some other software systems which have some intermediate-level capabilities include Stata (Computer Resources Center Inc., College Station TX), SPIDA (NHMRC Clinical Trials Centre, Eastwood, NSW Australia), and SAS (SAS Institute Inc., Cary NC).

9. CASE STUDY

Consider the 506-patient prostate cancer dataset from Byar and Green[67] which has also been analysed in references 68 and 69. The data are listed in reference 70, Table 46, and are available by Internet at utstat.toronto.edu in the directory /pub/data-collect. These data were from a randomized trial comparing four treatments for stage 3 and 4 prostate cancer, with almost equal numbers of patients on placebo and each of three doses of oestrogen. Four patients had missing values on all of the following variables: wt, pf, hx, sbp, dbp, ekg, hg, bm; two of these patients were also missing sz (see Table II for abbreviations). These patients will be excluded from consideration.

There are 354 deaths among the 502 patients. If we only wanted to test for a drug effect on survival time, a simple rank-based analysis would suffice. To be able to test for differential treatment effect or to estimate prognosis or expected absolute treatment benefit for individual

patients, however, we need a multivariable survival model.[3] First we consider fitting a full additive model which does not assume linearity of effect for any predictor. Categorical predictors will be expanded using dummy variables. For pf we could lump the last two categories since the last category has only two patients. Likewise, we could combine the last two levels of ekg. Continuous predictors will be expanded by fitting 4-knot restricted cubic spline functions, which contain two non-linear terms and thus have a total of 3 d.f. Table II defines the candidate predictors and lists their d.f. The variable stage is not listed as it can be predicted with high accuracy from sz, sg, ap, bm (stage could have been used as a predictor for imputing missing values on sz, sg).

There are a total of 36 candidate d.f. which should not be artificially reduced by 'univariable screening' or graphical assessments of association with death. This is about $\frac{1}{10}$ as many predictor d.f. as there are deaths, so there is some hope that a fitted model may validate. Let us also examine this issue by estimating the amount of shrinkage using equation (2). We use a Cox proportional hazards model for time until death. The UNIX S-Plus Design library fits the full model using restricted cubic spline expansions and makes use of Therneau's survival4 package in statlib[71] to perform the calculations. First we invoke the transcan function and impute functions (from statlib for any versions of S-Plus) to develop customized non-linear imputation equations for all predictors and to apply these equations to impute missing values.

```
# Define function for easy determination of whether a value is in a list
'%in%' ← function (a, b) match (a, b, nomatch = 0) > 0

levels(ekg) [levels(ekg) %in% c('old MI','recent MI')] ← 'MI'
# combines last 2 levels and uses a new name, MI

pf.coded  ← as.integer(pf)   # save original pf, re-code to 1-4
levels(pf) ← c(levels(pf) [1 : 3], levels(pf) [3])  # combine last 2 levels of original
w   ← transcan( ~ sz + sg + ap + sbp + dbp + age + wt + hg +
                  ekg + pf + bm + hx, imputed = T, impcat = 'tree')
sz  ← impute(w, sz)          # uses imputation rule w
sg  ← impute(w, sg)
age ← impute(w, age)
wt  ← impute(w, wt)
ekg ← impute(w, ekg)

dd  ← datadist(rx, age, wt, pf, pf.coded, heart, map, hg, sz, sg, ap, bm)
options(datadist = 'dd')       # datadist stores characteristics of raw data

units(dtime) ← 'Month'
S ← Surv(dtime, status! = 'alive')

f ← cph(S ~ rx + rcs(age,4) + rcs(wt,4) + pf + hx +
        rcs(sbp,4) + rcs(dbp,4) + ekg + rcs(hg,4) +
        rcs(sg,4) + rcs(sz,4) + rcs(ap,4) + bm)
```

The likelihood ratio χ^2 statistic is 140 with 36 d.f. This test is highly significant so some modelling is warranted. The AIC value (on the χ^2 scale) is $140 - 2 \times 36 = 68$. The rough shrinkage estimate is 0·743 (104/140) so we estimate that 26% of the model fitting will be noise, especially with regard to calibration accuracy. The approach of reference 2 is to fit this full model and to shrink predicted values. We will instead try to do data reduction (blinded to individual χ^2 statistics from the above model fit) to see if a reliable model can be obtained without shrinkage. A good approach at this point might be to perform a variable clustering analysis which for our purposes we will do informally. The data reduction strategy is listed in Table III. For ap, more exploration is desired to be able to model the shape of effect with such a highly skewed

Table III. Data reduction strategy (blinded to Y)

Variables	Reductions	d.f. saved
wt	Assume variable not important enough for 4 knots Use 3 knots	1
pf	Assume linearity	1
hx, ekg	Make new 0, 1, 2 variable and assume linearity: 2 = hx and ekg not normal and benign, 1 = either, 0 = none	5
sbp, dbp	Combine into mean arterial bp and use 3 knots: map = $\frac{2}{3}$ dpb + $\frac{1}{3}$ spb	4
sg	Use 3 knots	1
sz	Use 3 knots	1
ap	Look at shape of effect of ap in detail, and take log before expanding in spline to achieve numerical stability: add 2 knots	−2

distribution. Since we expect the tumour variables to be strong prognostic factors we will retain them as separate variables. No assumption will be made for the dose-response shape for oestrogen, as there was reason to expect a non-monotonic effect due to competing risks for cardiovascular death.

```
heart ← hx + I(ekg %in% c('normal','benign'))
label(heart) ← 'Heart Disease Code'
map ← (2*dbp + sbp)/3
label(map) ← 'Mean Arterial Pressure/10'

f ← cph(S ~ rx + rcs(age,4) + rcs(wt,3) + pf.coded +
        heart + rcs(map,3) + rcs(hg,4) +
        rcs(sg,3) + rcs(sz,3) + rcs(log(ap),6) + bm,
        x = T, y = T, surv = T, time.inc = 5*12)
# x, y for predict, validate, calibrate; surv, time.inc for calibrate
```

The total savings is thus 11 d.f. The likelihood ratio χ^2 is 126 with 25 d.f., with a slightly improved AIC of 76. The rough shrinkage estimate is slightly better at 0·80, but still worrisome. A further data reduction might be achieved by using the transcan transformations determined from self-consistency of predictors, but we will stop here and use this model.

Now assess this model in more detail by examining coefficients and summarizing multiple parameters within predictors using Wald statistics.

```
f      # writing an object name in S causes it to be printed

Cox Proportional Hazards Model

cph(formula = S ~ rx + rcs(age, 4) + rcs(wt, 3) + pf.coded + heart + rcs(map, 3) +
    rcs(hg, 4) + rcs(sz, 3) + rcs(sg, 3) + rcs(log(ap), 6) + bm,
    x = T, y = T, surv = T, time.inc = 5*12)

Obs Events Model L.R. d.f. P Score Score P R2
502    354   126  25  0  135    0  0·221

                          coef se(coef)          z        p
rx = 0·2 mg estrogen   3.74e-03 1.50e-01   0·0250   9·80e-01
rx = 1.0 mg estrogen  -4·21e-01 1·66e-01  -2·5427   1·10e-02
rx = 5.0 mg estrogen  -9·73e-02 1·58e-01  -0·6176   5·37e-01
                age   -1·17e-02 2·35e-02  -0·4995   6·17e-01
               age'    2·00e-02 3·86e-02   0·5190   6·04e-01
```

age''	2·71e − 01	4·95e − 01	0·5482	5·84e − 01
wt	−2·46e − 02	9·39e − 03	−2·6175	8·86e − 03
wt'	1·84e − 02	1·12e − 02	1·6379	1·01e − 01
pf.coded	2·25e − 01	1·21e − 01	1·8625	6·25e − 02
heart	4·18e − 01	8·08e − 02	5·1723	2·31e − 07
map	3·24e − 02	8·49e − 02	0·3817	7·03e − 01
map'	−4·57e − 02	9·41e − 02	−0·4857	6·27e − 01
hg	−1·56e − 01	7·68e − 02	−2·0343	4·19e − 02
hg'	7·42e − 02	2·10e − 01	0·3530	7·24e − 01
hg''	5·08e − 01	1·27e + 00	0·4014	6·88e − 01
sz	1·00e − 02	1·44e − 02	0·6955	4·87e − 01
sz'	8·79e − 03	2·37e − 02	0·3715	7·10e − 01
sg	7·19e − 02	7·86e − 02	0·9138	3·61e − 01
sg'	−7·04e − 03	9·83e − 02	−0·0716	9·43e − 01
ap	−7·96e − 01	3·11e − 01	−2·5584	1·05e − 02
ap'	4·89e + 01	2·18e + 01	2·2482	2·46e − 02
ap''	−3.64e + 02	1·59e + 02	−2·2909	2·20e − 02
ap'''	4·04e + 02	1·75e + 02	2·3067	2·11e − 02
ap''''	−9·69e + 01	4·16e + 01	−2·3311	1·97e − 02
bm	3·25e − 02	1·81e − 01	0·1790	8·58e − 01

```
# The terms with ', '', etc. after the name are cubic spline nonlinear terms
# The dose effect is apparently nonlinear.

anova(f)      # output was actually typesetted automatically using latex(anova(f))
              # latex requires the print.display package from statlib
```

There are 12 parameters associated with non-linear effects, and the overall test of linearity indicates the strong presence of non-linearity for at least one of the variables **age, wt, map, hg, sz, sg, ap** (see Table IV). There is a difference in survival time between at least two of the doses of oestrogen.

Now that we have a tentative model, let us examine the model's distributional assumptions. As mentioned in Section 4.3, the Schoenfeld partial residuals are an effective tool for checking the proportional hazards assumption in the Cox model. Grambsch and Therneau[72] have modified these residuals so that smoothed plots of them estimate the effect of predictors on the log instantaneous hazard rate as a function of follow-up time. Their scaled residuals estimate $\beta(t)$, the regression coefficient as a function of time. A messy detail is how to handle multiple regression coefficients per predictor. Here we do an approximate analysis in which each predictor is scored by adding up all the terms in the model to transform that predictor to be optimally related to the log hazard (at least if the *shape* of the effect does not change with time). In doing this we are temporarily ignoring the fact that the individual regression coefficients were estimated from the data. For dose of oestrogen, for example, we code the effect as 0 (placebo), 0·0037 (0·2 mg), −0·421 (1.0 mg), and −0·0973 (5.0 mg), and **age** is transformed as −0·0117 **age** + 0·02 **age'** + 0·271 **age''**, which in most simple form is

$$-1{\cdot}17\times10^{-2}\mathbf{age} + 3{\cdot}48\times10^{-5}(\mathbf{age}-56)_+^3 + 4{\cdot}71\times10^{-4}(\mathbf{age}-71)_+^3$$

$$-1{\cdot}01\times10^{-3}(\mathbf{age}-75)_+^3 + 5{\cdot}09\times10^{-4}(\mathbf{age}-80)_+^3$$

where $(x)_+$ means to ignore that term if $x \leqslant 0$, and the knots for age are 56, 71, 75 and 80 years.

In S-Plus the **predict** function easily summarizes multiple terms and produces a matrix (here, **z**) containing the total effects for each predictor. Matrix factors can easily be included in model

Table IV. Wald statistics for S

	χ^2	d.f.	P
rx	8·38	3	0·0387
age	12·85	3	0·0050
Non-linear	8·18	2	0·0168
wt	8·87	2	0·0118
Non-linear	2·68	1	0·1014
pf.coded	3·47	1	0·0625
heart	26·75	1	<0·0001
map	0·25	2	0·8803
Non-linear	0·24	1	0·6272
hg	11·85	3	0·0079
Non-linear	6·92	2	0·0314
sz	10·60	2	0·0050
Non-linear	0·14	1	0·7102
sg	3·14	2	0·2082
Non-linear	0·01	1	0·9429
ap	13·17	5	0·0218
Non-linear	12·93	4	0·0116
bm	0·03	1	0·8579
TOTAL NON-LINEAR	**30·28**	**12**	**0·0025**
TOTAL	**128·08**	**25**	**<0·0001**

formulae.

```
z ← predict(f, type = 'terms')          # required x = T above to store design
                                        # matrix
f.short ← cph(S ~ z, x = T, y = T)      # store x, y so can get residuals
```

The fit `f.short` based on the matrix `z` of single d.f. predictors has the same LR χ^2 of 126 as the fit f, but with a falsely low 11 d.f. All regression coefficients are unity.

Now get scaled Schoenfeld residuals separately for each predictor and test the proportional hazards assumption for each using the 'correlation with time' test. Also plot smoothed trends in the residuals. The `plot` method for `cox.zph` objects uses restricted cubic splines to smooth the relationship.

```
phtest ← cox.zph(f.short, transform = 'identity')
phtest
```

	rho	chisq	p
rx	0·12965	6·5451	0·0105
age	−0·08911	2.8518	0·0913
wt	−0·00878	0·0269	0·8697
pf.coded	−0·06238	1·4278	0·2321
heart	0·01017	0·0451	0·8319
map	0·03928	0·4998	0·4796
hg	−0·06678	1·7368	0·1876
sz	−0·05262	0·9834	0·3214
sg	−0·04276	0·6474	0·4210
ap	0·01237	0·0558	0·8133
bm	0·04891	0·9241	0·3364
GLOBAL	NA	15·3776	0·1659

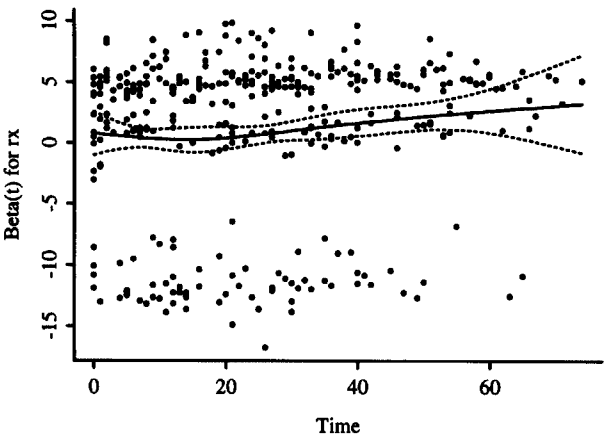


Figure 2. Raw and spline-smoothed scaled Schoenfeld residuals for dose of oestrogen, non-linearly coded from the Cox model fit, with ± 2 standard errors.[71]

Only the drug effect significantly changes over time ($P = 0{\cdot}01$ for testing the correlation rho between the scaled Schoenfeld residual and time), but when a global test of PH is done penalizing for 11 d.f., the P-value is 0·17. A graphical examination of the trends does not find anything interesting for the last 10 variables. A residual plot is drawn for rx alone and is shown in Figure 2.

```
plot(phtest, var = 'rx')
```

We will ignore the possible increase in effect of oestrogen over time. If this non–PH is real, a more accurate model might be obtained by stratifying on rx or by using a time × rx interaction as a time-dependent covariable.

Note that the model has several insignificant predictors. These will not be deleted, as that would not improve predictive accuracy and it would make confidence intervals for $\hat{\beta}$ or for predicted survival probabilities with the correct coverage probabilities hard to obtain.[64] At this point it would be reasonable to test pre–specified interactions. Here we will test all interactions with dose. Since the multiple terms for many of the predictors (and for rx) make for a great number of d.f. for testing interaction (and a loss of power), we will do approximate tests on the data-driven codings of predictors. P-values for these tests are likely to be somewhat anti-conservative.

```
z.dose ← z['rx']  # same as saying z[,1] – get first column
z.other ← z[,-1]   # all but the first column of z
f.ia ← cph(S ~ z.dose * z.other)
anova(f.ia)
```

Factor	Chi-Square	d.f.	P
z.dose (Factor + Higher Order Factors)	18·9	11	0·062
All Interactions	12·2	10	0·273
z.other (Factor + Higher Order Factors)	134·3	20	0·000
All Interactions	12·2	10	0·273
z.dose * z.other (Factor + Higher Order Factors)	12·2	10	0·273
TOTAL	137·3	21	0·000

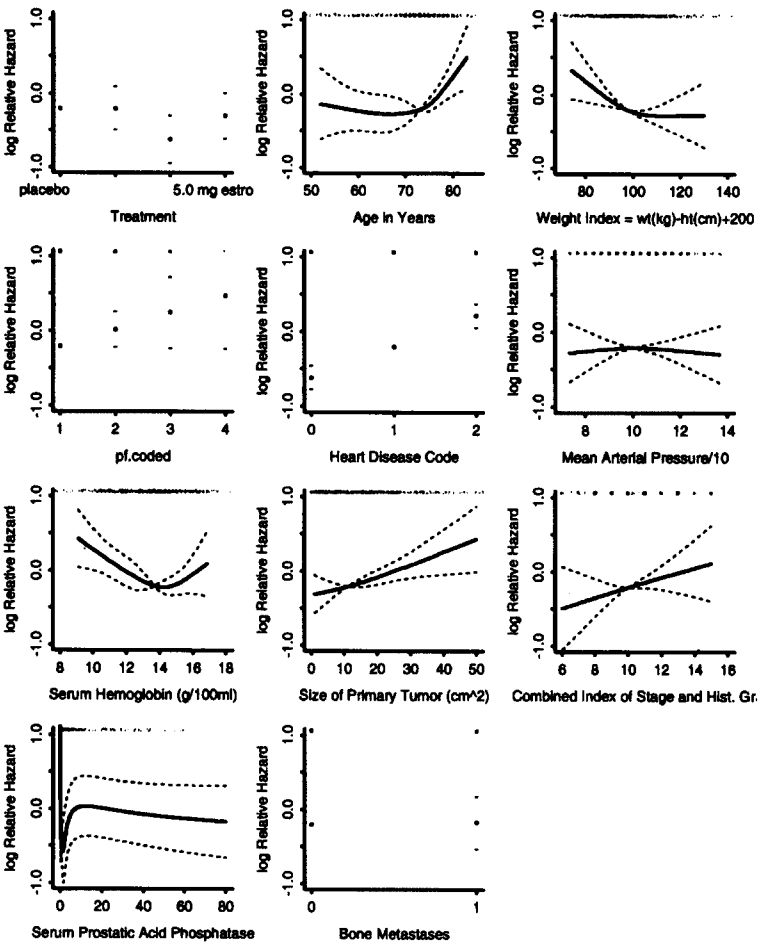


Figure 3. Shape of each predictor on log hazard of death. Y-axis shows $X\hat{\beta}$, but the predictors not plotted are set to reference values. 'Rug plots' on the top of each graph show the data density of the predictor. Note the highly non-monotonic relationship with ap, and the increased slope after age 70 which has been found in outcome models for various diseases

Here 'Factor + Higher Order Factors' means the combined main effect and interaction effect. The global test of additivity has $P = 0{\cdot}27$, so we will ignore the interactions (and also forget to penalize for having looked for them below!).

The following UNIX S-Plus statements plot how each predictor is related to the log hazard of death, along with 0·95 confidence bands. Note that due to a peculiarity of the Cox model the standard error of the predicted $X\hat{\beta}$ is zero at the reference values (medians here, for continuous predictors).

```
par(mfrow = c(3, 4))        # 4 x 3 matrix of graphs
r ← c(−1, 1)                # use common y-axis range for all
plot(f, rx = NA,            ylim = r)    NA → use default range for predictor
plot(f, age = NA,           ylim = r)
scatld(age)                              # scatld from statlib, for any S-Plus
plot(f, wt = NA,            ylim = r)    # scatld shows data density
...
```

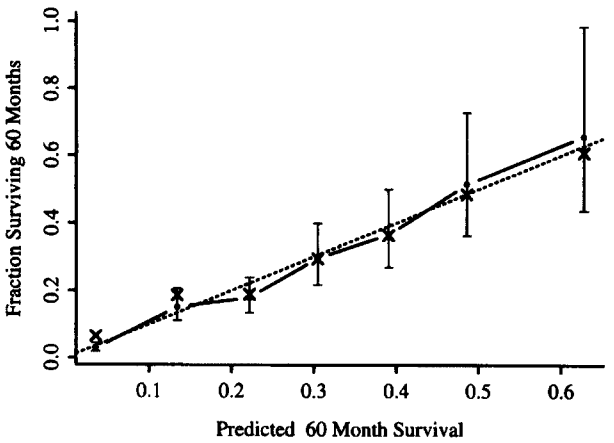


Figure 4. Bootstrap estimate of calibration accuracy for 5-year estimates from the final Cox model. Dots correspond to apparent predictive accuracy. × marks the bootstrap–corrected estimates

We first validate this model for Somers' D_{xy} rank correlation between predicted log hazard and observed survival time, and for slope shrinkage. The bootstrap is used (with 200 re-samples) to penalize for possible overfitting, as discussed in Section 6.

```
validate(f, B = 200, dxy = T, pr = T)
```

	index.orig	training	test	optimism	index. corrected	n
Dxy	−0·337377	−0·364644	−0·30976	−0·05488	−0·28250	200
R2	0·221444	0·261369	0·18445	0·07691	0·14453	200
Slope	1.000000	1.000000	0·78464	0·21536	0·78464	200

Here 'training' refers to accuracy when evaluated on the bootstrap sample used to fit the model, and 'test' refers to the accuracy when this model is applied without modification to the original sample. The apparent D_{xy} is −0·34, but a better estimate of how well the model will discriminate prognoses in the future is $D_{xy} = -0{\cdot}28$. The bootstrap estimate of slope shrinkage is 0·78, surprisingly close to the simple heuristic estimate. The shrinkage coefficient could easily be used to shrink predictions to yield better calibration.

Finally, we validate the model (without using the shrinkage coefficient) for calibration accuracy in predicting the probability of surviving 5 years. As detailed in Section 5, the bootstrap is used to estimate the optimism in how well predicted 5-year survival from the final Cox model tracks Kaplan–Meier 5-year estimates, stratifying by grouping patients in subsets with about 70 patients per interval of predicted 5-year survival.

```
plot(calibrate(f, B = 200, u = 5 * 12, m = 70))
```

The estimated calibration curves are shown in Figure 4. Bias–corrected calibration is very good except for the two groups with extremely bad prognosis – their survival is slightly better than predicted, consistent with regression to the mean. Even there, the absolute error is low despite a large relative error. Hence for this example it may not be worthwhile to develop a model using shrinkage.

Now compare this analysis with three previous analyses of this dataset. In all three analyses, all continuous covariables were arbitrarily categorized into intervals and scored with somewhat arbitrary category codes. In none of the three were **sbp, dbp, ekg, ap, bm** considered. Patients having missing values on any of the candidate predictors were excluded from consideration.

Turn first to Byar and Green,[67] who used an exponential survival model and dichotomized treatment by combining placebo and low dose and combining the two highest doses. The important predictors were found to be hx, sg, sz, hg, and the following interactions were detected in an exploratory analysis which did not control for multiple comparisons: rx × sg and rx × age. These interactions were not significant in the present model (even if dose were re-coded as in Byar and Green).

Kay[68] considered Cox models for various causes of death. For time until all-cause mortality, Kay found that the most important predictors were sz, hx, sg, age. The treatment along with age, hx were significant predictors of cardiovascular death. The treatment (in the opposite direction), and hg, sz, sg predicted cancer death. Treatment and age, wt predicted time until death from other causes.

Sauerbrei and Schumacher[69] also used a Cox model and an approach in which a backward elimination procedure was done for each of 100 bootstrap samples. The relative frequency of selection of variables as 'important' was used as the criterion for inclusion of variables in the final model. Variables were retained if they were selected $\geqslant 70$ times. All candidate predictors met this criterion. Treatment interactions involving age and sg were the most common interactions (56 and 48 bootstrap repetitions, respectively), but they did not meet the criterion for selection. The authors noted that these interactions were misleadingly more significant in a model which only adjusted for 'significant' predictors instead of all candidate predictors.

None of the three references just cited provided a model validation or quantified the predictive discrimination of the final model.

10. SUMMARY

Methods were described for developing clinical multivariable prognostic models and for assessing their calibration and discrimination. A detailed examination of model assumptions and an unbiased assessment of predictive accuracy will uncover problems that may make clinical prediction models misleading or invalid. The modelling strategy presented in Section 7 provides one sequence of steps for avoiding the pitfalls of multivariable modelling so that its many advantages can be realized.

ACKNOWLEDGEMENTS

This work was supported by research grants HL-17670, HL-29436, HL-36587, HL-45702 and HL-09315 from the National Heart, Lung and Blood Institute, Bethesda, Maryland, research grants HS-03834, HS-05635, HS-06503, HS-06830, and HS-07137 from the Agency for Health Care Policy and Research, Rockville, Maryland, and grants from the Robert Wood Johnson Foundation, Princeton, NJ.

REFERENCES

1. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. and Rosati, R. A., 'Evaluating the yield of medical tests', *Journal of the American Medical Association*, **247**, 2543–2546 (1982).
2. Spiegelhalter, D. J. 'Probabilistic prediction in patient management', *Statistics in Medicine*, **5**, 421–433 (1986).
3. Knaus, W. A., Harrell, F. E., Fisher, C. J., Wagner, D. P., Opan, S. M., Sadoff, J. C., Draper, E. A., Walawander, C. A., Conboy, K. and Grasela, T. H. 'The clinical evaluation of new drugs for sepsis: A prospective study design based on survival analysis', *Journal of the American Medical Association*, **270**, 1233–1241 (1993).

4. Donner, A. 'The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values', *American Statistician*, **36**, 378–381 (1982).
5. Roberts, J. S. and Capalbo, G. M. 'A SAS macro for estimating missing values in multivariate data', *Proceedings of the Twelfth Annual SAS Users Group International Conference*, (Cary NC), SAS Institute, 939–941 (1987).
6. Buck, S. F. 'A method of estimation of missing values in multivariate data suitable for use with an electronic computer', *Journal of the Royal Statistical Society, Series B*, **22**, 302–307 (1960).
7. Timm, N. H. 'The estimation of variance-covariance and correlation matrices from incomplete data', *Psychometrika*, **35**, 417–437 (1970).
8. Kuhfeld, W. F. 'The PRINQUAL procedure', in *SAS/STAT User's Guide*, 4th edn., vol. **2**, SAS Institute, Cary NC, 1990, chapter 34, pp. 1265–1323.
9. Schemper, M. 'Non-parametric analysis of treatment-covariate interaction in the presence of censoring', *Statistics in Medicine*, **7**, 1257–1266 (1988).
10. Cox, D. R. 'The regression analysis of binary sequences (with discussion)', *Journal of the Royal Statistical Society, Series B*, **20**, 215–242 (1958).
11. Walker, S. H. and Duncan, D. B. 'Estimation of the probability of an event as a function of several independent variables', *Biometrika*, **54**, 167–178 (1967).
12. van Houwelingen J. C. and le Cessie, S. 'Logistic regression, a review', *Statistica Neerlandica*, **42**, 215–232 (1988).
13. Collett, D. *Modelling Binary Data*. Chapman and Hall, London, 1991.
14. Cox, D. R. 'Regression models and life-tables (with discussion)', *Journal of the Royal Statistical Society, Series B*, **34**, 187–220 (1972).
15. Collett, D. *Modelling Survival Data in Medical Research*. Chapman and Hall, London 1994.
16. Lawless, J. F. *Statistical Models and Methods for Lifetime Data*. Wiley, New York 1982.
17. Derksen S. and Keselman, H. J. 'Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables', *British Journal of Mathematical and Statistical Psychology*, **45**, 265–282 (1992).
18. Grambsch, P. M. and O'Brien, P. C. 'The effects of transformations and preliminary tests for non-linearity in regression', *Statistics in Medicine*, **10**, 697–709 (1991).
19. Verweij, P. and van Houwelingen, H. C. 'Penalized likelihood in Cox regression', *Statistics in Medicine*, **13**, 2427–2436 (1994).
20. Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B. and Rosati, R. A. 'Regression modelling strategies for improved prognostic prediction', *Statistics in Medicine*, **3**, 143–152 (1984).
21. Marshall, G., Grover, F. L., Henderson, W. G. and Hammermeister, K. E. 'Assessment of predictive models for binary outcomes: an empirical approach using operative death from cardiac surgery', *Statistics in Medicine*, **13**, 1501–1511 (1994).
22. Jolliffe, I. T. *Principal Component Analysis*, Springer-Verlag, New York, 1986.
23. Jackson, J. E. *A User's Guide to Principal Components*, Wiley, New York, 1991.
24. Smith, L. R., Harrell, F. E. and Muhlbaier, L. H. 'Problems and potentials in modelling survival', in: Grady, M. L. and Schwartz, H. A. (eds.), *Medical Effectiveness Research Data Methods (Summary Report)*, *AHCPR Pub. No. 92–0056* US Dept. of Health and Human Services, Agency for Health Care Policy and Research, Rockville, Maryland, 1992, pp. 151–159.
25. Durrleman, S. and Simon, R. 'Flexible regression models with cubic splines', *Statistics in Medicine*, **8**, 551–561 (1989).
26. Harrell, F. E., Lee, K. L. and Pollock, B. G. 'Regression models in clinical studies: determining relationships between predictors and response', *Journal of the National Cancer Institute*, **80**, 1198–1202 (1988).
27. Sleeper, L. A. and Harrington, D. P. 'Regression splines in the Cox model with application to covariate effects in liver disease', *Journal of the American Statistical Association*, **85**, 941–949 (1990).
28. Landwehr, J. M., Pregibon, D. and Shoemaker, A. C. 'Graphical methods for assessing logistic regression models (with discussion)', *Journal of the American Statistical Association*, **79**, 61–83 (1984).
29. Hosmer, D. W. and Lemeshow, S. *Applied Logistic Regression*. Wiley, New York, 1989.
30. Therneau, T. M., Grambsch, P. M. and Fleming, T. R. 'Martingale-based residuals for survival models', *Biometrika*, **77**, 216–218 (1990).
31. Schoenfeld, D. 'Partial residuals for the proportional hazards regression model', *Biometrika*, **69**, 239–241 (1982).

32. Pettitt A. N. and Bin Daud, I. 'Investigating time dependence in Cox's proportional hazards model', *Applied Statistics*, **39**, 313–329 (1990).
33. Harrell, F. E., Pollock, B. G. and Lee, K. L. 'Graphical methods for the analysis of survival data', in *Proceedings of the Twelfth Annual SAS Users Group International Conference*, Cary, NC, pp. 1107–1115, SAS Institute, Inc., 1987.
34. van Houwelingen, J. C. and le Cessie, S. 'Predictive value of statistical models', *Statistics in Medicine*, **8**, 1303–1325 (1990).
35. Friedman, J. H. 'A variable span smoother', Technical Report 5, Laboratory for Computational Statistics, Department of Statistics, Stanford University, 1984.
36. Cleveland, W. S. 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association*, **74**, 829–836 (1979).
37. Statistical Sciences, *S-Plus User's Manual, Version 3.2.*, StatSci, a division of MathSoft, Inc., Seattle WA, 1993.
38. Brier, G. W. 'Verification of forecasts expressed in terms of probability,' *Monthly Weather Review*, **75**, 1–3 (1950).
39. Kaplan, E. L. and Meier, P. 'Nonparametric estimation from incomplete observations', *Journal of the American Statistical Association*, **53**, 457–481 (1958).
40. Copas, J. B. 'Regression, prediction and shrinkage (with discussion)', *Journal of the Royal Statistical Society, Series B*, **45**, 311–354 (1983).
41. Copas, J. B. 'Cross–validation shrinkage of regression predictors', *Journal of the Royal Statistical Society, Series B*, **49**, 175–183 (1987).
42. Gray, R. J. 'Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis', *Journal of the American Statistical Association*, **87**, 942–951 (1992).
43. Goodman, L. A. and Kruskal, W. H. *Measures of Association for Cross-Classifications*, Springer-Verlag, New York 1979.
44. Brown, B. W., Hollander, M. and Korwar, R. M. 'Nonparametric tests of independence for censored data, with applications to heart transplant studies', in: Proschan, F. and Serfling,R. J. (eds), *Reliability and Biometry*, SIAM, Philadelphia, 1974.
45. Schemper, M. 'Analyses of associations with censored data by generalized Mantel and Breslow tests and generalized Kendall correlation', *Biometrical Journal*, **26**, 309–318 (1984).
46. Bamber, D. 'The area above the ordinal dominance graph and the area below the receiver operating characteristic graph', *Journal of Mathematical Psychology*, **12**, 387–415 (1975).
47. Hanley, J. A. and McNeil, B. J. 'The meaning and use of the area under a receiver operating characteristic (ROC) curve', *Radiology*, **143**, 29–36 (1982).
48. Liu, K. and Dyer, A. R. 'A rank statistic for assessing the amount of variation explained by risk factors in epidemiologic studies', *American Journal of Epidemiology*, **109**, 597–606 (1979).
49. Nagelkerke, N. J. D. 'A note on a general definition of the coefficient of determination', *Biometrika*, **78**, 691–692 (1991).
50. Lee, K. L., Pryor, D. B., Harrell, F. E., Califf, R. M., Behar, V. S., Floyd, W. L., Morris, J. J., Waugh, R. A., Whalen, R. E. and Rosati, R. A. 'Predicting outcome in coronary disease: Statistical models versus expert clinicians', *American Journal of Medicine*, **80**, 553–560 (1986).
51. Harrell, F. E. and Lee, K. L. 'A comparison of the *discrimination* of discriminant analysis and logistic regression under multivariate normality', in Sen, P. K. (ed), *Biostatistics: Statistics in Biomedical, Public Health, and Environmental Sciences. The Bernard G. Greenberg Volume*, North-Holland, New York, 1985, pp. 333–343.
52. Korn, E. L. and Simon, R. 'Measures of explained variation for survival data', *Statistics in Medicine*, **9**, 487–503 (1990).
53. Picard, R. R. and Berk, K. N. 'Data splitting', *American Statistician*, **44**, 140–147 (1990).
54. Efron, B. 'Estimating the error rate of a prediction rule: improvement on cross-validation', *Journal of the American Statistical Association*, **78**, 316–331 (1983).
55. Linnet, K. 'Assessing diagnostic tests by a strictly proper scoring rule', *Statistics in Medicine*, **8**, 609–618 (1989).
56. Efron, B. and Gong, G. 'A leisurely look at the bootstrap, the jackknife, and cross-validation', *American Statistician*, **37**, 36–48 (1983).
57. Efron, B. 'How biased is the apparent error rate of a prediction rule?', *Journal of the American Statistical Association*, **81**, 461–470 (1986).
58. Efron, B. and Tibshirani, R. *An Introduction to the Bootstrap.* Chapman and Hall, New York, 1993.

59. Roecker, E. B. 'Prediction error and its estimation for subset-selected models', *Technometrics*, **33**, 459–468 (1991).
60. Breiman, L. 'The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error', *Journal of the American Statistical Association*, **87**, 738–754 (1992).
61. Atkinson, A. C. 'A note on the generalized information criterion for choice of a model', *Biometrika*, **67**, 413–418 (1980).
62. Crawford, S. L., Tennstedt, S. L. and McKinlay, J. B. 'A comparison of analytic methods for non-random missingness of outcome data', *Journal of Clinical Epidemiology*, **48**, 209–219 (1995).
63. Mantel, N. 'Why stepdown procedures in variable selection', *Technometrics*, **12**, 621–625 (1970).
64. Altman, D. G. and Andersen, P. K. 'Bootstrap investigation of the stability of a Cox regression model', *Statistics in Medicine*, **8**, 771–783 (1989).
65. Hurvich, C. M. and Tsai, C. L. 'The impact of model selection on inference in linear regression', *American Statistician*, **44**, 214–217 (1990).
66. Harrell, F. E. 'Design: S-Plus functions for biostatistical/epidemiologic modelling, testing, estimation, validation, graphics, prediction, and typesetting by storing enhanced model design attributes in the fit. Programs available from statlib@lib.stat.cmu.edu. Send E–mail 'send design from S", 1994.
67. Byar, D. P. and Green, S. B. 'The choice of treatment for cancer patients based on covariate information: application to prostate cancer', *Bulletin Cancer*, Paris, **67**, 477–488 (1980).
68. Kay, R. 'Treatment effects in competing-risks analysis of prostate cancer data', *Biometrics*, **42**, 203–211 (1986).
69. Sauerbrei, W. and Schumacher, M. 'A bootstrap resampling procedure for model building: Application to the Cox regression model', *Statistics in Medicine*, **11**, 2093–2109, (1992).
70. Andrews, D. F. and Herzberg, A. M. *Data*. New York, Springer-Verlag, 1985.
71. Therneau, T. 'Survival4: S functions for survival analysis. Programs available from statlib@lib.stat.cmu.edu. Send E–mail 'send survival4 from S,' 1995.
72. Grambsch, P. and Therneau, T. 'Proportional hazards tests and diagnostics based on weighted residuals', *Biometrika*, **81**, 515–526 (1994).

TUTORIAL IN BIOSTATISTICS

DEVELOPMENT OF A CLINICAL PREDICTION MODEL FOR AN ORDINAL OUTCOME:
The World Health Organization Multicentre Study of Clinical Signs and Etiological Agents of Pneumonia, Sepsis and Meningitis in Young Infants

FRANK E. HARRELL, Jr.[1]*, PETER A. MARGOLIS[2], SANDY GOVE[3], KAREN E. MASON[3], E. KIM MULHOLLAND[4], DEBORAH LEHMANN[5], LULU MUHE[6], SALVACION GATCHALIAN[7] AND HEINZ F. EICHENWALD[8]
and the
WHO/ARI YOUNG INFANT MULTICENTRE STUDY GROUP

[1] *Division of Biostatistics and Epidemiology, Department of Health Evaluation Sciences, University of Virginia, Charlottesville, U.S.A.*
[2] *The Division of Community Pediatrics, University of North Carolina, Chapel Hill, U.S.A.*
[3] *Programme for the Control of Acute Respiratory Infection (ARI) of the World Health Organization, Geneva, Switzerland*
[4] *MRC, The Gambia*
[5] *Papua New Guinea Institute of Medical Research, Goroka*
[6] *Department of Paediatrics and Child Health, Addis Ababa University, Ethiopia*
[7] *Research Institute for Tropical Medicine, Alabang, Philippines*
[8] *Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, U.S.A.*

SUMMARY

This paper describes the methodologies used to develop a prediction model to assist health workers in developing countries in facing one of the most difficult health problems in all parts of the world: the presentation of an acutely ill young infant. Statistical approaches for developing the clinical prediction model faced at least two major difficulties. First, the number of predictor variables, especially clinical signs and symptoms, is very large, necessitating the use of data reduction techniques that are blinded to the outcome. Second, there is no uniquely accepted continuous outcome measure or final binary diagnostic criterion. For example, the diagnosis of neonatal sepsis is ill-defined. Clinical decision makers must identify infants likely to have positive cultures as well as to grade the severity of illness. In the WHO/ARI Young Infant Multicentre Study we have found an ordinal outcome scale made up of a mixture of laboratory and diagnostic markers to have several clinical advantages as well as to increase the power of tests for risk factors. Such a mixed ordinal scale does present statistical challenges because it may violate constant slope

* Correspondence to: Frank E. Harrell Jr, PhD, Box 600, Health Sciences Center, University of Virginia, Charlottesville VA 22908, U.S.A. E-mail: fharrell@virginia.edu

Contract grant sponsor: Agency for Health Care Policy and Research; contract grant number: HS-06830, HS-07137

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
 ISBN: 0-470-02365-1

assumptions of ordinal regression models. In this paper we develop and validate an ordinal predictive model after choosing a data reduction technique. We show how ordinality of the outcome is checked against each predictor. We describe new but simple techniques for graphically examining residuals from ordinal logistic models to detect problems with variable transformations as well as to detect non-proportional odds and other lack of fit. We examine an alternative type of ordinal logistic model, the continuation ratio model, to determine if it provides a better fit. We find that it does not but that this model is easily modified to allow the regression coefficients to vary with cut-offs of the response variable. Complex terms in this extended model are penalized to allow only as much complexity as the data will support. We approximate the extended continuation ratio model with a model with fewer terms to allow us to draw a nomogram for obtaining various predictions. The model is validated for calibration and discrimination using the bootstrap. We apply much of the modelling strategy described in Harrell, Lee and Mark (*Statist. Med.* **15**, 361–387 (1998)) for survival analysis, adapting it to ordinal logistic regression and further emphasizing penalized maximum likelihood estimation and data reduction.

1. INTRODUCTION

The presentation of an acutely ill young infant presents health workers in all parts of the world with one of their most difficult problems. Serious infections are the main cause of morbidity and mortality in infants under 3 months of age in developing countries. Diagnosis is difficult – meningitis or pneumonia might appear as clear clinical syndromes, but more often the picture is mixed and the infant is labelled as 'sepsis'. Even in industrialized countries, treatment is usually based on clinical impressions supported by laboratory data, which by itself is often inconclusive. In developing countries, clinical signs are the only tools available in most places. The ability to detect serious bacterial infection early in young infants is important in defining appropriate prevention and treatment strategies. A better clinical prediction rule to be used by peripheral health workers might result in more appropriate referral to hospital as well as less antibiotic use in very low-risk infants.

We set out to determine which combination of clinical signs most accurately predict the group of infants who have meningitis, sepsis or pneumonia. There is no single 'gold standard' against which to correlate these signs. It is tempting to use as an endpoint the physician's expert clinical diagnosis. This would induce a circularity which would inflate the predictive discrimination of the prediction model, because clinical signs are major determinants of the overall clinical impression. Death is the only endpoint that can be ascertained for every infant, but truly ill infants who were successfully treated early with antibiotics may not die.

Even without a gold standard, however, there are a number of generally agreed laboratory tests which could be used to construct a reasonable outcome scale, including cerebro-spinal fluid (CSF) culture from a lumbar puncture (LP), blood culture (BC), arterial oxygen saturation (SaO_2, a measure of lung function), and chest X-ray (CXR). In this study, 249 infants died, and death could be placed at the top of an ordinal outcome scale. Many of the deaths were related to starting treatment too late so they were not preventable with antibiotics. For this paper, we choose to ignore death (but not to exclude patients who died) when constructing the scale as the main goal was to predict treatable disease. Ignoring death resulted in some of the clinical signs being stronger predictors. Two-thirds of the deaths were redistributed to other positive outcome categories.

We model an ordinal outcome scale using the proportional odds (PO) form of an ordinal logistic model[1] and the forward continuation ratio (CR) ordinal logistic model.[2] (see references 3–18 for some excellent background references, applications, and extensions to the ordinal

models.) We predicted this ordinal outcome using clinical symptoms, signs, and basic variables such as age, weight, temperature and respiratory rate. Nine major statistical problems had to be addressed in these analyses:

1. How does one avoid estimating a separate coefficient for the large number of clinical signs? (Overfitting and poor model validation would result if all signs were treated as separate candidate variables in the model.)[19]
2. Can expert clinicians assign weights for signs *a priori* that adequately predict the outcomes?
3. Given that the clinical signs can be combined in meaningful ways, how should a cluster of such variables be quantified? Should weights for multiple signs be summed or should the cluster be scored using the weight associated with the most severe sign present? Is the union of all signs (that is, presence of any sign) within a cluster an adequate summary of that cluster?
4. How can continuous predictors such as respiratory rate or temperature be modelled flexibly without assuming linearity?
5. Since the response variable is a hierarchical assignment made up of disparate measurements is the proportional odds assumption likely to be violated?
6. Will another type of ordinal logistic model provide a better fit?
7. How can the constant slopes assumption be relaxed without causing overfitting?
8. How does one diagram the final ordinal model so that field health workers can quickly obtain predicted risks of various severities of outcome?
9. How does one validate an ordinal regression model without sacrificing sample size?

Section 2 gives a brief overview of the World Health Organization/Acute Respiratory Infection (WHO/ARI) Multicentre Study design. Section 3 provides the definition of the ordinal outcome scale. In Section 4 we discuss how clinical signs were scored (individually) and clustered. Section 5 tests the adequacy of weights specified by subject-matter specialists and depicts the utility of various scoring schemes using a tentative ordinal logistic model. Section 6 depicts a simple way to assess the assumption of ordinality of the response with respect to each predictor, and to examine the PO and CR assumptions separately for each predictor. In section 7 we derive a tentative proportional odds model using cluster scores and using regression splines to allow other predictors to be flexibly related to the log odds of an outcome. Section 8 shows how residuals from binary logistic models can be adapted to the ordinal case, and uses smoothed residual plots to assess the proportional odds assumption with respect to each predictor. Section 9 examines the fit of a continuation ratio model. Section 10 shows how the CR model can easily be extended (and fitted using standard software) to allow some or all of the regression coefficients to vary with cut-offs of the response level as well as to provide formal tests of constant slopes. Section 11 shows how penalized maximum likelihood estimation is used to improve future predictive accuracy. In Section 12 the full model is approximated by a sub-model, and a nomogram is constructed for the approximate model. Section 13 demonstrates how the ordinal model is validated using the bootstrap. Many of the methods discussed here were discussed in Harrell *et al.*[20] where the focus was on survival analysis. This modelling strategy used here generally follows that paper, with additional stress on penalized estimation.

Cole *et al.*[17] also presented a case study in developing a PO ordinal logistic model for diagnosing illness in infants under 6 months of age. In their study, which was based on patients who were less severely ill than those in the present study, the ordinal outcome was physicians' subjective assessment of the severity of illness. The analysis was based on stepwise variable selection of individual clinical signs which would be expected to prevent the model calibration to

be accurate for very low and very high risk infants. That paper included a nice example of optimally rounding regression coefficients so that a simple severity score could be derived.

For almost all steps of the analysis, computer code is shown, both to make the steps more concrete as well as to show their feasibility. All analyses were done using S-plus version 3.2[21] on UNIX using Sun Sparcstation 2 and 10 computers in conjunction with the `Design` library of UNIX and Microsoft Windows S-plus functions.[22] For binary and PO logistic models `Design` has a general penalized maximum likelihood estimation facility in the `lrm` function. It also has a function `cr.setup` which allows the CR model to be fitted in an extremely flexible way using a binary logistic model on a modified input data set. `Design` is available at `http://www.med.virginia.edu/medicine/clinical/hes/biostat.htm`. `varclus`, `transcan`, `impute`, and `scatld` are separate functions in the `Hmisc` library in `statlib`, also written by the first author.

2. STUDY DESIGN

The WHO/ARI Multicentre Study of Clinical Signs and Etiological Agents of Pneumonia, Sepsis and Meningitis in Young Infants was undertaken in Ethiopia, The Gambia, Papua New Guinea, and the Phillippines to collect data that would allow better screening criteria to be derived for finding infants at high risk of serious infection.[23] Standardized laboratory and clinical evaluations (clinical history, risk factors such as low birth weight, CXR, SaO_2, BC, LP etc.) were done. Infants brought for primary care were enrolled if *any* of the following were present: cough; difficult, fast or noise breathing; fever or hypothermia; not feeding well (less than half of normal intake); abnormally sleepy or difficult to wake; convulsions; rectal temperature $\geqslant 37{\cdot}5°C$ or $\leqslant 35{\cdot}5°C$; or mother volunteered that the baby was very sick, irritable, or has stopped breathing or turned blue/black. Infants were excluded if the illness began in hospital (except for delivery), the clinic visit was for trauma, burn, or routine care such as immunization, weight < 1500 g during the first 48 hours of life, there was a documented episode of previous pneumonia, sepsis, or meningitis within the last 3 weeks, if an obvious congenital malformation was present, or if the infant had previously been enrolled in the study. 8418 infants were screened and the 4552 having a positive symptom without an exclusion were enrolled.

To be a candidate for BC and CXR, an infant had to have a clinical indication for one of the three diseases, according to prespecified criteria in the study protocol ($n = 2398$). Blood work-up (but not LP) and CXR was also done on a random sample intended to be 10 per cent of infants having no signs or symptoms suggestive of infection ($n = 175$).* Infants with signs suggestive of meningitis had LP. All 4552 infants received a full physical exam and standardized pulse oximetry to measure SaO_2. The vast majority of infants getting CXR had the X-rays interpreted by three independent radiologists.

For determining outcomes, BC was considered positive if the blood culture was definite or probable for bacterial infection.[23] CSF was positive if the CSF culture was positive or there were more than 10 white cells in the CSF, after a correction was made for a bloody tap. CXR is positive if all radiologists who read the CXR and ruled it interpretable classified the infant's X-ray as definitely or probably abnormal.

* Some mothers refused having blood drawn for their children, adding a slight bias to the remaining sampled group which made them slightly more sick. At one site, the sample was systematic (every fifth patient) but sampling was done more often when mothers refused to participate

Table I. Ordinal outcome scale

Outcome level Y	Definition	n	Fraction in outcome level		
		4552	BC, CXR indicated ($n = 2398$)	Not indicated ($n = 1979$)	Random sample* ($n = 175$)
0	None of the above	3551	0·63	0·96	0·91
1	$90\% \leqslant SaO_2 < 95\%$ or CXR +	490	0·17	0·04†	0·05
2	BC + or CSF + or $SaO_2 < 90\%$	511	0·21	0·00‡	0·03

* A separate sample of patients not indicated for laboratory work-up but having it anyway
† SaO_2 was measured but CXR was not done
‡ *Assumed* zero since neither BC nor LP were done

The analyses which follow are not corrected for verification bias[24] with respect to BC, LP and CXR, but Section 3 has some data describing the extent of the problem.

3. ORDINAL OUTCOME SCALE

Rationale and details of the outcome scale construction are found in a background paper.[25] As discussed by Follman,[26] it is useful to derive new outcome variables (risk scores) by observing how non-fatal events predict death. We followed this scheme but extended it with a two-stage strategy. First, we found that the more important non-fatal response measures, BC+, CSF+, and severe hypoxemia ($SaO_2 < 90$ per cent, altitude adjusted), had roughly equal weight in predicting death* so the union of these findings was used as the top outcome level. Next, CXR and moderate hypoxemia ($90 < SaO_2 < 95$ per cent) were examined for their association with the probability that the patient had BC+ or CSF+. These two markers were found to have the same associations with this worse outcome (probabilities of BC+ $\cup$ CSF+ were equal for CXR+ and for $SaO_2 \in [90$ per cent, 95 per cent), and were lower but equal for CXR− and $SaO_2 \geqslant 95$ per cent).

Patients were then assigned to the worst qualifying outcome category. Table I shows the definition of the ordinal outcome variable Y and shows the distribution of Y by the laboratory work-up strategy.

The effect of verification bias is a false negative fraction of 0·03 for $Y = 2$, from comparing the detection fraction of zero for $Y = 2$ in the 'not indicated' group with the observed positive fraction of 0·03 in the random sample that was fully worked up. The extent of verification bias in $Y = 1$ is $0·05 - 0·04 = 0·01$. In what follows, these biases will be ignored.

4. VARIABLE CLUSTERING

Expert clinical judgement was used to enumerate a list of clinical variables to collect, including 47 clinical signs. The list reflects the content of an expert paediatric examination. As a first step in

* Proportion of deaths for BC+, BC−, CSF+, CSF−, $SaO_2 < 90$ per cent, $SaO_2 \geqslant 90$ per cent were, respectively, 0·30, 0·08, 0·29, 0·05, 0·25, 0·04

coding the predictor variables, all questionnaire items that were connected (for example, using skip rules such as 'if condition was present, what was its severity?') were scored as a single variable using equally spaced codes, with 0–3 representing, for example, sign not present, mild, moderate, severe. The resulting list of clinical signs with their abbreviations is given in Table II. The signs are organized into clusters as will be discussed below. Here, hx stands for history, ausc for auscultation, and hxprob for history of problems. Two signs (qcr, hcm) were listed twice because they were later placed into two clusters each.

When there are many candidate predictors, several authors[20,27–29] have demonstrated that because variable clustering reduces the number of regression coefficients to test or estimate, it results in better validating models than either stepwise modelling or fitting a full model with at least one regression coefficient per candidate predictor. A commonly used variable clustering technique is a rotation of principal components (see Section 3.2 of D'Agostino *et al.*,[29] Chapters 5, 6, 8, 9, 12, 14 of Cureton and D'Agostino[30] and Sarle[31]) by which variables are separated into groups so that the first principal component of that group of variables explains the majority (for example 0·8) of the variance for that group of variables and so that the correlation of individual variables in different groups is low. Instead of using specialized variable clustering procedures, there are advantages to using traditional cluster analysis to cluster variables. Traditional cluster analysis uses a distance matrix to cluster subjects, but cluster analysis can also cluster variables by using a 'similarity matrix' as input. The advantages of this clustering approach are: (i) a multitude of similarity measures are available, including ones that allow for non-monotonic relationships,[32] (ii) there are many clustering techniques available (a concise summary may be found in Venables and Ripley[33], p. 311–315), including some that allow overlap between clusters. Here we used the matrix of squared Spearman rank correlation coefficients in the similarity matrix. The varclus function in the Hmisc library, which uses the S-plus hierarchical clustering function hclust, was used as follows:

```
vclust ← varclus(~ illd + hlt + slpm + slpl + wake + convul + hfa +
                   hfb + hfe  + hap  + hcl  + hcm + hcs  + hdi + fde +
                   chi + twb  + ldy  + apn  + lcw + nfl  + str + gru +
                   coh + ccy  + jau  + omph + csd + csa  + aro + qcr +
                   con + att  + mvm  + afe  + absu + stu  + deh + dcp +
                   crs + abb  + abk  + whz  + hdb + smi2 + abd + conj +
                   oto + puskin,  sim="spearman")
plot(vclust)
```

The output appears in Figure 1.

Overall, the statistical clusterings made clinical sense, for example, stu (skin turgor) and deh (dehydrated) are closely related, and these two are somewhat less related to abk (sunken fontanelle) than to hdi (history of diarrhoea). In many cases, the clusters suggested by such output can be used directly in data reduction and summary scale construction. More often, though, the output serves as a starting point for clinicians to use in constructing more meaningful clinical clusters. That was the case here, and the clusters in Table II were the consensus of the clinicians who were the investigators in the WHO/ARI study. Prior subject matter knowledge plays a key role at this stage in the analysis.

See Sections 3 and 4 of D'Agostino *et al.*[29] for a full description of a process for eliciting clusters from subject-matter experts and then using statistical clustering techniques to check that each cluster represents only one central concept.

Table II. Clinical signs

Cluster name	Sign abbreviation	Name of sign	Values
bul.conv	abb	bulging fontanelle	0–1
	convul	hx convulsion	0–1
hydration	abk	sunken fontanelle	0–1
	hdi	hx diarrhoea	0–1
	deh	dehydrated	0–2
	stu	skin turgor	0–2
	dcp	digital capillary refill	0–2
drowsy	hcl	less activity	0–1
	qcr	quality of crying	0–2
	csd	drowsy state	0–2
	slpm	sleeping more	0–1
	wake	wakes less easy	0–1
	aro	arousal	0–2
	mvm	amount of movement	0–2
agitated	hcm	crying more	0–1
	slpl	sleeping less	0–1
	con	consolability	0–2
	csa	agitated state	0–1
crying	hcm	crying more	0–1
	hcs	crying less	0–1
	qcr	quality of crying	0–2
	smi2	smiling ability × age > 42 days	0–2
reffort	nfl	nasal flaring	0–3
	lcw	lower chest in-drawing	0–3
	gru	grunting	0–2
	ccy	central cyanosis	0–1
stop.breath	hap	hx stop breathing	0–1
	apn	apnoea	0–1
ausc	whz	wheezing	0–1
	coh	cough heard	0–1
	crs	crepitation	0–2
hxprob	hfb	fast breathing	0–1
	hdb	difficulty breathing	0–1
	hlt	mother report respiratory problems	none, chest, other
feeding	hfa	hx abnormal feeding	0–3
	absu	sucking ability	0–2
	afe	drinking ability	0–2
labor	chi	previous child died	0–1
	fde	fever at delivery	0–1
	ldy	days in labour	1–9
	twb	water broke	0–1
abdominal	abd	abdominal distension	0–4
	jau	jaundice	0–1
	omph	omphalitis	0–1
fever.ill	illd	ge-adjusted number days ill	
	hfe	hx fever	0–1
pustular	conj	conjuctivitis	0–1
	oto	otoscopy impression	0–2
	puskin	pustular skin rash	0–1

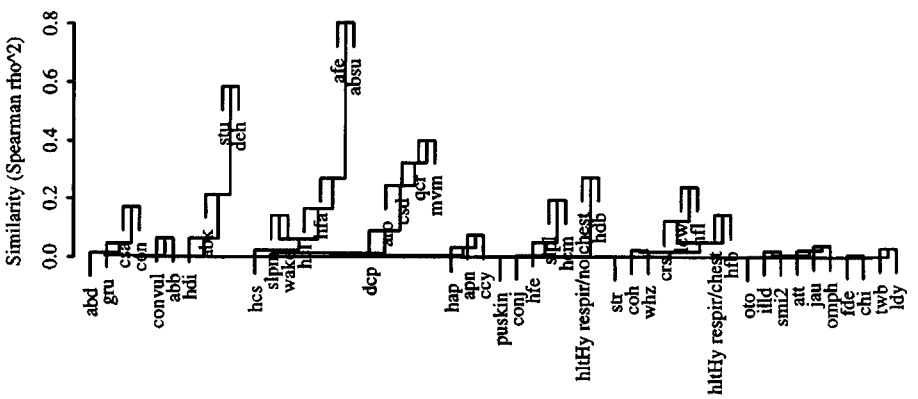


Figure 1. Hierarchical variable clustering using Spearman ρ^2 as a similarity measure for all pairs of variables. Note that because the hlt variable was nominal, it is represented by two dummy variables here

5. THE PROPORTIONAL ODDS MODEL AND DEVELOPING CLUSTER SUMMARY SCORES

The clusters listed in Table II were first scored by the first principal component – the linear combination of signs (using the codes in Table II) that explains the maximum variance of all signs in a cluster explainable by a single dimension (a single linear combination).[29,34,35] We denote the first principal component by PC_1. Knowing that the resulting weights may be too complex for clinical use, the primary reasons for analysing the principal components was to see if some of the clusters could be removed from consideration so that the clinicians would not spend time developing scoring rules for them. We decided to 'peak' at Y to assist in scoring clusters at this point, but to do so in a very structured way that did not involve the examination of a large number of individual coefficients.

We did not actually compute PC_1s on the raw signs but rather used a psychometric scaling technique similar to that of Kuhfeld[36] which is implemented in the S-plus transcan function. Here, for any cluster which contains a sign with more than two levels, the levels are automatically re-scored so as to increase the variance explained by the PC_1. This could be called a non-linear principal components analysis.

To judge any cluster scoring scheme, we had to pick a tentative outcome model. For this purpose we chose the most commonly used ordinal logistic model, which was described in Walker and Duncan[1] and later called the *proportional odds (PO) model* by McCullagh.[14] The PO model is stated as follows:

$$\Pr[Y \geqslant j | X] = \frac{1}{1 + \exp[-(\alpha_j + X\beta)]} \tag{1}$$

where there is an implicit assumption that the regression coefficients (β) are independent of j, the cut-off level for Y. Note that the PO model makes no assumption whatever about the magnitude of spacings between levels of Y. By using the 14 PC_1s corresponding to the 14 clusters, the fitted PO model had a likelihood ratio (LR) χ^2 of 1155 with 14 d.f., and the predictive discrimination of

Table III. Clinician combinations, rankings and scorings of signs

Cluster	Combined/ranked signs in order of severity	Weights
bul.conv	abb∪convul	0–1
drowsy	hcl, qcr > 0, csd > 0∪slpm∪wake, aro > 0, mvm > 0	0–5
agitated	hcm, slpl, con = 1, csa, con = 2	0, 1, 2, 7, 8, 10
refort	nfl > 0, lcw > 1, gru = 1, gru = 2, ccy	0–5
ausc	whz, coh, crs > 0	0–3
feeding	hfa = 1, hfa = 2, hfa = 3, absu = 1∪afe = 1, absu = 2∪afe = 2	0–5
abdominal	jau∪abd > 0∪omph	0–1

the clusters was quantified by a Somers' D_{xy} rank correlation between $X\hat{\beta}$ and Y of 0·596.* The following clusters were not statistically important predictors and we assumed that the lack of importance of the PC_1s in predicting Y (adjusted for the other PC_1s) justified a conclusion that no sign within that cluster was clinically important in predicting Y: hydration, hxprob, pustular, crying, fever.ill, stop.breath, labor. This list was identified using a backward step-down procedure on the full model. The total Wald χ^2 for these 7 PC_1s was 22·4 ($P = 0{\cdot}002$). The reduced model had LR $\chi^2 = 1133$ with 7 d.f., $D_{xy} = 0{\cdot}591$. The bootstrap validation in Section 13 penalized for fitting the 7 predictors.

During a meeting of the study group, the clinicians were asked to rank the clinical severity of signs within each potentially important cluster. During this step, the clinicians also ranked severity levels of some of the component signs, and some cluster scores were simplified, especially when the signs within a cluster occurred infrequently. The clinicians also assessed whether the severity points or weights should be equally spaced, assigning unequally spaced weights for one cluster (agitated). The resulting rankings and sign combinations are shown in Table III. The signs or sign combinations separated by a comma are treated as separate categories, whereas some signs were unioned ('or'-ed) when the clinicians deemed them equally important. As an example, if an additive cluster score was to be used for drowsy, the scorings would be 0 = none present, 1 = hcl, 2 = qcr > 0, 3 = csd > 0 or slpm or wake, 4 = aro > 0, 5 = mvm > 0 and the scores would be added.

This table reflects some data reduction already (unioning some signs and selection of levels of ordinal signs) but more reduction is needed. Even after signs are ranked within a cluster, there are various ways of assigning the cluster scores. We investigated six methods. We started with the purely statistical approach of using PC_1 to summarize each cluster. Second, all sign combinations within a cluster were unioned to represent 0–1 cluster score. Third, only sign combinations thought by the clinicians to be severe were unioned, resulting in drowsy=aro>0 or mvm>0, agitated=csa or con=2, reffort=lcw>1 or gru>0 or ccy, ausc=crs>0, and feeding=absu>0 or afe>0. For clusters that are not scored 0–1 in Table III, the fourth summarization method was a hierarchical one which used the weight of the worst applicable category as the cluster score. For example, if aro=1 but mvm=0, drowsy would be scored as 4. The fifth method counted the number of positive signs in the cluster. The sixth method summed the weights of all signs or sign

* See reference 20 for details; $D_{xy} = 2(c - \frac{1}{2})$ where c is the probability of concordance between pairs of $X\hat{\beta}$ and Y values, which is a generalization of a receiver operating characteristic curve area

Table IV. Predictive information of various cluster scoring strategies

Scoring method	LR χ^2	d.f.	AIC
PC_1 of each cluster	1133	7	1119
Union of all signs	1045	7	1031
Union of higher categories	1123	7	1109
Hierarchical (worst sign)	1194	7	1180
Additive, equal weights	1155	7	1141
Additive using clinician weights	1183	7	1169
Hierarchical, data-driven weights	1227	25	1177

combinations present. Finally, the worst sign combination present was again used as in the second method, but the points assigned to the category were data driven ones obtained by using extra dummy variables. This provides an assessment of the adequacy of the clinician-specified weights. By comparing rows 4 and 7 in Table IV we see that response data-driven sign weights have a slightly worse Akaike information criterion (AIC) or LR $\chi^2 - 2 \times$ d.f. (which penalizes the model for complexity[37]), indicating that the number of extra β parameters estimated was not justified by the improvement in χ^2. The hierarchical method, using the clinicians' weights, performed quite well. The only cluster with inadequate clinician weights was ausc – see following. The PC_1 method, without any guidance, performed well, as in Harrell *et al.*[19] The only reasons not to use it are that it requires a coefficient for every sign in the cluster and coefficients are not translatable into simple scores such as 0, 1, ...

Representation of clusters by a simple union of selected signs or of all signs is inadequate, but otherwise the choice of methods is not very important in terms of explaining variation in Y. We chose the fourth method, a hierarchical severity point assignment (using weights which were prespecified by the clinicians), for its ease of use and of handling missing component variables (in most cases) and potential for speeding up the clinical exam (examining to detect more important signs first). Because of what was learned regarding the relationship between ausc and Y, we modified the ausc cluster score by redefining it as ausc = crs > 0 (crepitations present). Note that neither the 'tweaking' of ausc nor the examination of the seven scoring methods displayed in Table IV will be taken into account in the model validation.

One attractive alternative approach that we did not try was the battery reduction strategy described in Chapter 12 of Cureton and D'Agostino[30] in which one finds a subset of the variables in each cluster whose PC_1 adequately represents the whole cluster's PC_1.

6. ASSESSING ORDINALITY OF Y FOR EACH X, AND UNADJUSTED CHECKING OF PO AND CR ASSUMPTIONS

A basic assumption of all commonly used ordinal regression models is that the response variable behaves in an ordinal fashion with respect to each predictor. Assuming that a predictor X is linearly related to the log odds of some appropriate event, a simple way to check for ordinality is to plot the mean of X stratified by levels of Y (denote these by $\hat{E}(X \mid Y = y)$). These means should be in a consistent order. If for many of the Xs, two adjacent categories of Y do not distinguish the means, that is evidence that those levels of Y should be pooled.

One can also estimate the mean or expected value of $X|Y=j$ given that the ordinal model assumption hold. This is a useful tool for examining those assumptions, at least in an unadjusted fashion. For simplicity, assume that X is discrete, and let $P_{jx} = \Pr(Y=j|X=x, \text{model})$ be the probability that $Y=j$ given $X=x$ that is dictated from the model being fitted, with X being the only predictor in the model. Then

$$\Pr(X=x|X=j, \text{model}) = \Pr(Y=j|X=x, \text{model})\Pr(X=x)/\Pr(Y=j)$$
$$E(X|Y=j, \text{model}) = \sum_x xP_{jx}\Pr(X=x)/\Pr(Y=j) \tag{2}$$

and the expectation can be estimated by

$$\hat{E}(X|Y=j, \text{model}) = \sum_x x\hat{P}_{jx}f_x/g_j \tag{3}$$

where $\hat{P}_{jx}$ denotes the estimate of P_{jx} from the fitted 1-predictor model*, f_x is the frequency of $X=x$ in the sample of size n, and g_j is the frequency of $Y=j$ in the sample. This estimate can be computed conveniently without grouping the data by X. For n subjects let the n values of X be $x_1, x_2, \ldots, x_n$. Then

$$\hat{E}(X|Y=j) = \sum_{i=1}^{n} x_i\hat{P}_{jx_i}/g_j. \tag{4}$$

Figure 2 was produced by the S-plus function plot.xmean.ordinaly in the Design library, which plots simple Y-stratified means overlaid with $\hat{E}(X|Y=j, \text{model})$, with j on the x-axis. Here we expect strongly non-linear effects for temp, rr and hrat, so for those predictors we plot the mean absolute differences from suitable 'normal' values as an approximate solution:

```
par(mfrow=c(3,4)) #3 x 4 matrix of plots
plot.xmean.ordinaly(Y ~ age + abs(temp-37) + abs(rr-60) + abs(hrat-125) +
                    waz + bul.conv + drowsy + agitated + reffort +
                    ausc + feeding + abdominal, cr=T)
```

The plot is shown in Figure 2. Y does not seem to operate in an ordinal fashion with respect to age, |rr-60| or ausc. For the other variables, ordinality holds, and PO holds reasonably well except for bul.conv, drowsy and abdominal. For heart rate, the PO assumption appears to be satisfied perfectly. CR model assumptions appear to be no less tenuous than PO assumptions, at least when fitting one variable at a time.

There is a relationship between score residuals defined later in equation (6) and a slightly different comparison between stratified means of X and expected values under the model. Suppose that simple means and expected values were computed under the condition that $Y \geq j$ instead of the condition $Y=j$. Then $\hat{E}(X|Y \geq j) - \hat{E}(X|Y \geq j, \text{model})$ is proportional to the mean of the score residuals for the PO model.

7. A TENTATIVE FULL PROPORTIONAL ODDS MODEL

Using summary cluster scores that were developed in Section 5, the original list of 14 clusters with 47 signs was reduced to 7 predictors as listed in Table III: two unions of signs (bul.conv,

* For inner values of Y in the PO models, these probabilities are differences between terms given by equation (1)

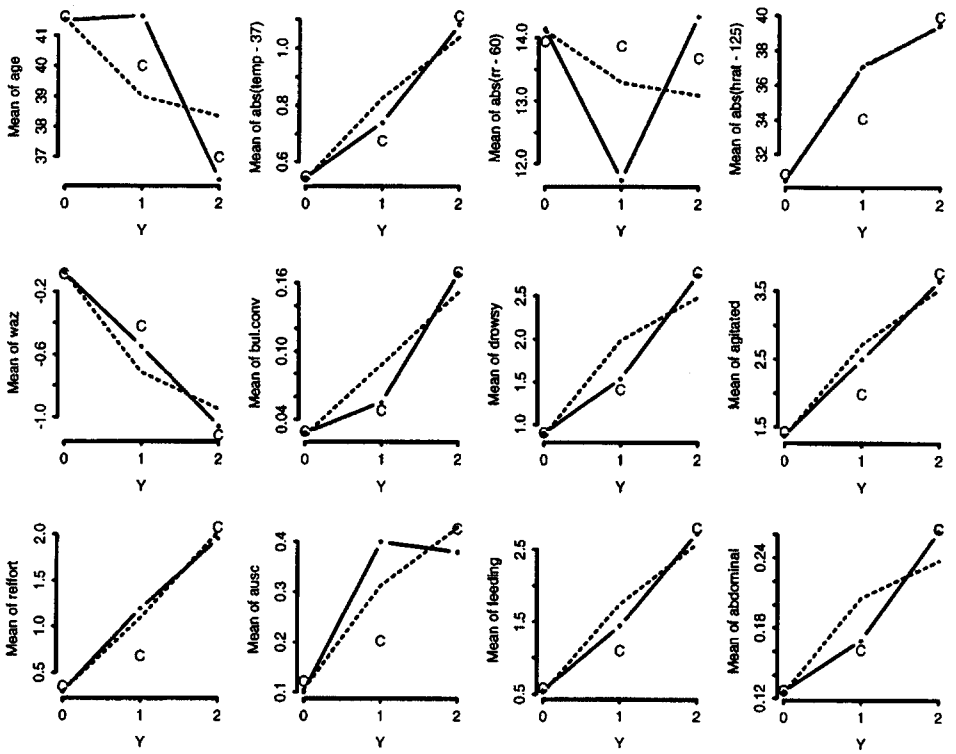


Figure 2. Examination of the ordinality of Y for each predictor by assessing if differing Y distinguish the mean X and if the the trend is monotonic. Solid lines connect the simple stratified means, and dashed lines connect the estimated expected value of $X|Y = j$ given that PO holds. Estimated expected values from the CR model are marked with c's

abdominal); one single sign (ausc), and four 'worst category' point assignments (drowsy, agitated, reffort, feeding). Seven clusters were dropped because of weak associations with Y*. Such a limited use of variable selection avoids most of the severe problems inherent with that technique such as the lack of replication of the list of 'significant' variables if one sampled repeatedly from the same population.[20]

At this point in model development we added to the model age and vital signs: temp (temperature); rr (respiratory rate); hrat (heart rate); and waz, weight-for-age Z-score. Since age was expected to modify the interpretation of temp, rr and hrat, and interactions between continuous variables would be difficult to use in the field, we categorized age into three intervals: 0–6 days ($n = 302$); 7–59 days ($n = 3042$); and 60–90 days ($n = 1208$).† This was done with the S-plus cut2 function from the Hmisc library in statlib:

```
ageg ← cut2(age, c(7,60))
```

* These clusters were reinstated as candidate predictors in the final model validation, to penalize for having tested them for association with Y

† These age intervals were also found to adequately capture more of the interaction effects

The new variables temp, rr, hrat, waz were missing in, respectively, $n = 13$, 11, 147 and 20 infants. Because the three vital sign variables are somewhat correlated with each other, customized imputation models were developed to impute all the missing values without assuming linearity or even monotonicity of any of the regressions. The S-plus transcan and impute functions from Hmisc were used to impute vital signs as follows:

```
vsign.trans ← transcan(~ temp + hrat + rr, imputed=T)
temp ← impute(vsign.trans, temp)
hrat ← impute(vsign.trans, hrat)
rr   ← impute(vsign.trans, rr)
```

After transcan estimated optimal restricted cubic spline transformations, temp could be predicted with adjusted $R^2 = 0{\cdot}17$ from hrat and rr, hrat could be predicted with adjusted $R^2 = 0{\cdot}14$ from temp and rr, and rr could be predicted with adjusted R^2 of only 0·06. The first two R^2, while not large, mean that customized imputations are more efficient than imputing with constants. Imputations on rr were closer to the median rr of 48/minute as compared with the other two vital signs whose imputed values have more variation. In a similar manner, waz was imputed using age, birth weight, head circumference, body length, and prematurity (adjusted R^2 for predicting waz from the others was 0.74).

The continuous predictors temp, hrat, rr were not assumed to linearly relate to the log odds that $Y \geqslant j$. Flexible piecewise cubic polynomials (restricted cubic spline functions[38–41] with 5 knots or join points for temp, rr and 4 knots for hrat,waz*) were used to model the effects of these variables, using the rcs function with the binary and PO logistic regression function lrm in the Design library:

```
f1 ← lrm(Y ~ ageg * (rcs(temp,5) + rcs(rr,5) + rcs(hrat,4)) + rcs(waz,4) +
         bul.conv + drowsy + agitated + reffort + ausc +
         feeding + abdominal, x=T, y=T) #x=T, y=T used by resid( ) below
```

Here the asterisk in the formula indicates that main effects and interactions are to be fitted. This model has LR χ^2 of 1393 with 45 d.f. and $D_{xy} = 0{\cdot}653$. Wald tests of non-linearity and interaction are obtained using the statement anova(f1), whose output is shown in Table V.†

The bottom four lines of the table are the most important. First, there is strong evidence that some associations with Y exist (45 d.f. test) and very strong evidence of non-linearity in one of the vital signs or in waz (26 d.f. test). There is moderately strong evidence for an interaction effect somewhere in the model (22 d.f. test). The anova output does not contain Wald χ^2 statistics for main effects alone as these are meaningless; main effect parameters are pooled with interaction ('higher order') parameters to yield meaningful overall tests for predictors. We see that the grouped age variable ageg is predictive of Y, but mainly as an effect modifier for rr. temp is extremely non-linear and rr is moderately so. hrat, a difficult variable to measure reliably in young infants, is perhaps not important enough ($\chi^2 = 19{\cdot}0$, 9 d.f.) to keep in the final model.

* Four knots were used for hrat because it was thought to be less important *a priori*, and fewer were used for waz because it was thought to operate almost linearly

† Actually, the statement which produced this output is latex(anova(f1)), which typeset the output using the LaTeX document processing language[42]

Table V. Wald statistics for Y in the proportional odds model

	LR χ^2	d.f.	P
ageg (factor + higher order factors)	41·49	24	0·0147
All interactions	40·48	22	0·0095
temp (factor + higher order factors)	37·08	12	0·0002
All interactions	6·77	8	0·5617
Non-linear (factor + higher order factors)	31·08	9	0·0003
rr (factor + higher order factors)	81·16	12	< 0·0001
All interactions	27·37	8	0·0006
Non-linear (factor + higher order factors)	27·36	9	0·0012
hrat (factor + higher order factors)	19·00	9	0·0252
All interactions	8·83	6	0·1836
Non-linear (factor + higher order factors)	7·35	6	0·2901
waz	35·82	3	< 0·0001
Non-linear	13·21	2	0·0014
bul.conv	12·16	1	0·0005
drowsy	17.79	1	< 0·0001
agitated	8·25	1	0·0041
reffort	63·39	1	< 0·0001
ausc	105·82	1	< 0·0001
feeding	30·38	1	< 0·0001
abdominal	0·74	1	0·3895
ageg × temp (factor + higher order factors)	6·77	8	0·5617
Non-linear	6·40	6	0·3801
Non-linear interaction: $f(A, B)$ versus AB	6·40	6	0·3801
ageg × rr (factor + higher order factors)	27·37	8	0·0006
Non-linear	14·85	6	0·0214
Non-linear interaction: $f(A, B)$ versus AB	14.85	6	0·0214
ageg × hrat (factor + higher order factors)	8·83	6	0·1836
Non-linear	2·42	4	0·6587
Non-linear interaction: $f(A, B)$ versus AB	2·42	4	0·6587
Total non-linear	78·20	26	< 0·0001
Total interaction	40·48	22	0·0095
Total non-linear + interaction	96·31	32	< 0·0001
Total	1073·78	45	< 0·0001

The clinicians did not presuppose that any of the clinical signs had special importance in combination with other signs or with vital signs. Therefore interactions involving the clinical signs, which would have been great in number, were not examined.

8. RESIDUALS FOR CHECKING THE PROPORTIONAL ODDS ASSUMPTION

Peterson and Harrell[15] developed score and likelihood ratio tests for testing the PO assumption. The score test is used in the SAS LOGISTIC procedure,[43] but it yields P-values that are far too small in many cases.[15] Other techniques, especially graphical ones, are needed for verifying PO. Schoenfeld residuals[44] are very effective[45] in checking the proportional hazards assumption in the Cox[46] survival model. For the PO model one could analogously compute each subject's contribution to the first derivative of the log-likelihood function with respect to β_m, average them

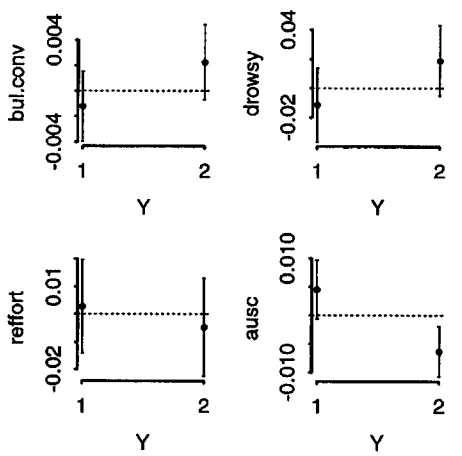


Figure 3. Binary logistic model score residuals for binary events derived from two cut-offs of the ordinal response Y. Note that the mean residuals, marked with closed circles, correspond closely with differences between solid and dashed lines at $Y = 1, 2$ in Figure 2. Bars are 0·95 confidence limits. Score residual assessments for spline-expanded variables such as rr would have required one plot per d.f.

separately by levels of Y, and examine trends in the residual plots. A few examples have shown that such plots are usually hard to interpret. Easily interpreted score residual plots for the PO model can be constructed however by using the fitted PO model to predict a series of binary events $Y \geqslant j, j = 1, 2, \ldots, k$, using the corresponding predicted probabilities

$$\hat{P}_{ij} = \frac{1}{1 + \exp[-(\hat{\alpha}_j + X_i\hat{\beta})]} \tag{5}$$

where X_i stands for a vector of predictors for subject i. Then, after forming an indicator variable for the event currently being predicted ($[Y_i \geqslant j]$), one computes the score (first derivative) components U_{im} from an ordinary binary logistic model:

$$U_{im} = X_{im}([Y_i \geqslant j] - \hat{P}_{ij}) \tag{6}$$

for the subject i and predictor m. Then, for each column of U, plot the mean $\bar{U}_{\cdot m}$ and confidence limits, with Y (that is, j) on the x-axis. For each predictor the trend against j should be flat if PO holds.*

For the tentative PO model, score residuals for four of the variables were plotted using

```
par(mfow=c(2,2))
resid(f1, 'score.binary', pl=T, which=c(17,18,20,21))
```

The result is shown in Figure 3. We see strong evidence of non-PO for ausc and moderate evidence for drowsy and bul.conv, in agreement with Figure 2.

* If $\hat{\beta}$ were derived from separate binary fits, all $\bar{U}_{\cdot m} \equiv 0$

In binary logistic regression, *partial residuals* are very useful because after the analyst fits linear effects for all predictors, computes partial residuals, and smooths the relationship between each predictor and its partial residuals, the resulting trend is an estimate of the true relationship between each predictor and the log odds. The partial residual is defined as follows, for the ith subject and mth predictor variable:[47,48]

$$r_{im} = \hat{\beta}_m X_{im} + \frac{Y_i - \hat{P}_i}{\hat{P}_i(1 - \hat{P}_i)} \tag{7}$$

where

$$\hat{P}_i = \frac{1}{1 + \exp[-(\hat{\alpha} + X_i\hat{\beta})]}. \tag{8}$$

A smoothed plot (for example, using the moving linear regression algorithm in lowess[49]) of X_{im} versus r_{im} provides a non-parametric estimate of how X_m relates to the log relative odds that $Y = 1 | X_m$.

For ordinal Y, we just need to compute binary model partial residuals for all cut-offs j:

$$r_{im} = \hat{\beta}_m X_{im} + \frac{[Y_i \geqslant j] - \hat{P}_{ij}}{\hat{P}_{ij}(1 - \hat{P}_{ij})} \tag{9}$$

then to make a plot for each m showing smoothed partial residual curves for all j, looking for similar shapes and slopes for a given predictor for all j. Each curve provides an estimate of how X_m relates to the relative log odds that $Y \geqslant j$. Since partial residuals allow examination of predictor transformation (linearity) while simultaneously allowing examination of PO (parallelism), partial residual plots are generally preferred over score residual plots for ordinal models.

In Figure 4, smoothed partial residual plots were obtained for all predictors, after first fitting a simple model in which every predictor was assumed to operate linearly. Interactions were temporarily ignored and age was used as a continuous variable:

```
f2 ← lrm(Y ~ age + temp + rr + hrat + waz +
         bul.conv + drowsy + agitated + reffort + ausc +
         feeding + abdominal, x=T, y=T)
par(mfrow=c(3,4))

resid(f2, 'partial', pl=T)   # pl=T : plot
```

The degree of non-parallelism generally agreed with the degree of non-flatness in Figure 3 and with the other score residual plots which were not shown. The partial residuals show that temp is highly non-linear and that it is much more useful in predicting $Y = 2$. For the cluster scores, the linearity assumption appears reasonable, except possibly for drowsy. Other non-linear effects will be taken into account using splines as before (except for age, which will be categorized).

A model can have significant lack of fit with respect to some of the predictors and still yield quite accurate predictions. To see if the case for this PO model, we computed predicted probabilities of $Y = 2$ for all infants from the model and compared these with predictions from a binary logistic model derived specifically to predict $\Pr(Y = 2)$ (that is, a model with no assumptions connecting different levels of Y). The mean absolute difference in predicted probabilities between the two models is only 0·02, but the 0·90 quantile of that difference is 0·059. For

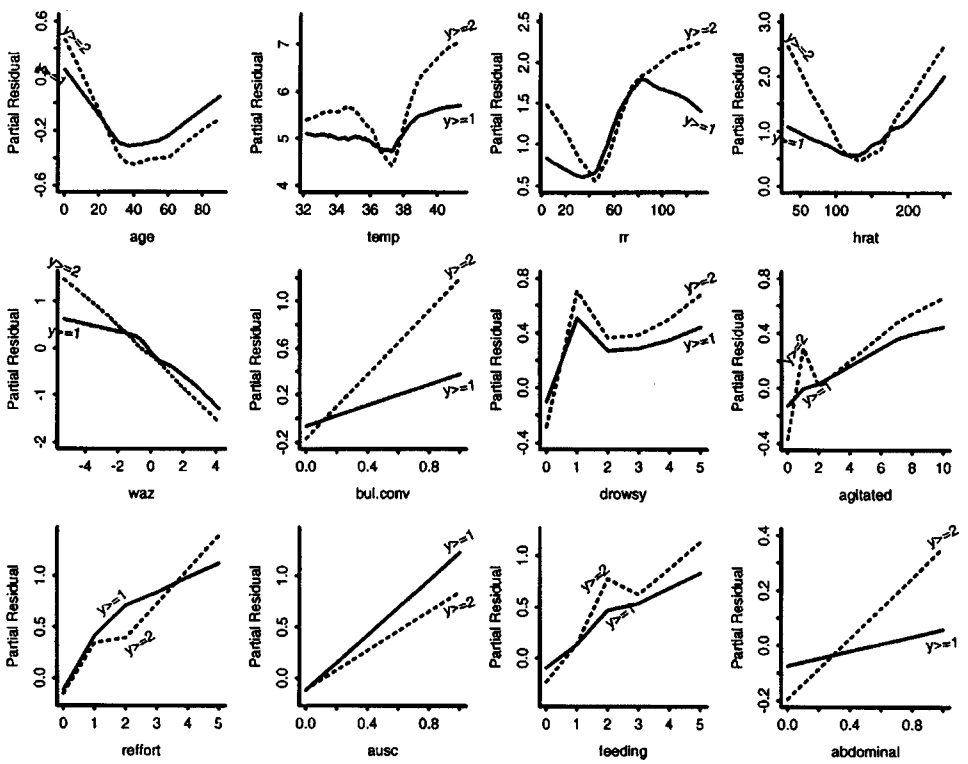


Figure 4. Smoothed partial residuals corresponding to two cut-offs of Y, from a model in which all predictors were assumed to operate linearly and additively. The smoothed curves estimate the actual predictor transformations needed.

high-risk infants, discrepancies of 0·2 were common. Therefore we elected to consider a different model.

9. CONTINUATION RATIO ORDINAL LOGISTIC MODEL

Unlike the PO model, which is based on cumulative probabilities, the continuation ratio (CR) model is based on conditional probabilities. The (forward) CR model[2,6,8] is stated as follows for $Y = 0, \ldots, k$ (here $k = 2$):

$$\begin{aligned} \Pr(Y = j \mid Y \geqslant j, X) &= \frac{1}{1 + \exp[-(\theta_j + X\gamma)]} \\ \text{logit}(Y = 0 \mid Y \geqslant 0, X) &= \text{logit}(Y = 0 \mid X) \\ &= \theta_0 + X\gamma \\ \text{logit}(Y = 1 \mid Y \geqslant 1, X) &= \theta_1 + X\gamma. \end{aligned} \tag{10}$$

The CR model has been said to be likely to fit ordinal responses when subjects have to 'pass through' one category to get to the next (which is the case here with respect to SaO_2). The CR model is a discrete version of the Cox proportional hazards model.

To check CR model assumptions, binary logistic model partial residuals are again valuable. We fit a sequence of binary logistic models using a series of binary events and the corresponding applicable (increasingly small) subsets of subjects, and plot smoothed partial residuals against X for all of the binary events. In S-plus we now fit the sequence of binary fits and then use the plot.lrm.partial function, which assembles partial residuals for a sequence of fits and constructs one graph per predictor:

```
cr0 ← lrm(Y==0 ~ age + temp + rr + hrat + waz +
          bul.conv + drowsy + agitated + reffort + ausc +
          feeding + abdominal, x=T, y=T)
# Use the update function to save repeating model right hand side
# An indicator variable for Y=1 is the response variable below
cr1 ← update(cr0, Y==1 ~ ., subset=Y>=1)
plot.lrm.partial(cr0, cr1, center=T)
```

The output is in Figure 5. There is not much more parallelism here than in Figure 4. For the two most important predictors, ausc and rr, there are strongly differing effects for the differing events being predicted (for example, $Y = 0$ vs. $Y = 1 | Y \geqslant 1$). As is often the case, there is no one constant β model that satisfies assumptions with respect to all predictors simultaneously, especially when there is evidence for non-ordinality for ausc in Figure 2. The CR model will need to be generalized to adequately fit this data set.

10. EXTENDED CONTINUATION RATIO MODEL

By comparing Figures 4 and 5 it is seen that the CR model in its ordinary form has no advantage over the PO model for this data set. The PO model has been extended by Peterson and Harrell[15] to allow for unequal slopes for some or all of the X's for some or all levels of Y. This partial PO model requires specialized software, and with the demise of SAS Version 5 PROC LOGIST, software is not currently available. Armstrong and Sloan[6] and Berridge and Whitehead[8] showed how the CR model can be fitted using ordinary binary logistic model software, after certain rows of the X matrix are duplicated and a new binary Y vector is constructed. For each subject, one constructs separate records by considering successive conditions $Y \geqslant 0$, $Y \geqslant 1$, ... , $Y \geqslant k-1$ for a response variable with values $0, 1, \ldots, k$. The binary response for each applicable condition or 'cohort' is set to 1 if the subject failed at the current 'cohort' or 'risk set', that is, if $Y = j$ where the cohort being considered is $Y \geqslant j$. The constructed cohort variable is carried along with the new X and Y. This variable is considered to be categorical and its coefficients are fitted by adding $k-1$ dummy variables to the binary logistic model. The CR model is restated as follows:

$$\Pr(Y = j | Y \geqslant j, X) = \frac{1}{1 + \exp[-(\alpha + \theta_j + X\gamma)]}. \tag{11}$$

Here α is an overall intercept, $\theta_0 \equiv 0$, and $\theta_1, \ldots, \theta_{k-1}$ are increments from α. In S-plus notation, the model is

$$\text{y} \sim \text{cohort} + \text{X1} + \text{X2} + \text{X3} + \ldots,$$

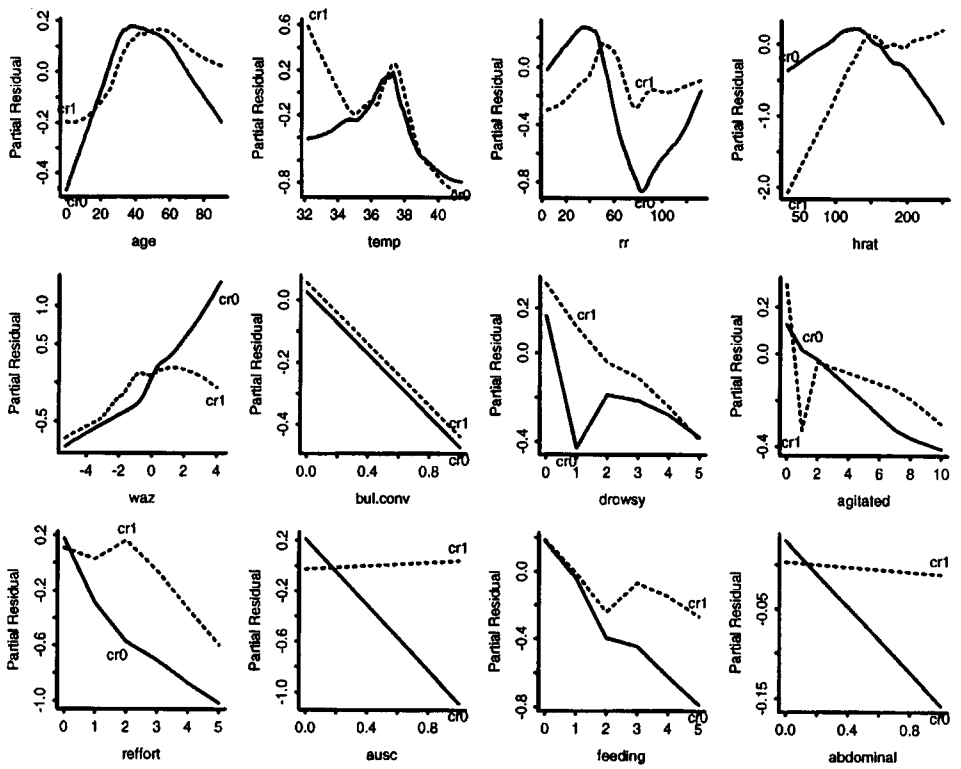


Figure 5. lowess smoothed partial residual plots for binary models which are components of an ordinal continuation ratio model

with `cohort` denoting a polytomous variable and the columns of X denoted by `X1, X2` ... , etc. The CR model can be extended to allow for some or all of the γ's to change with the cohort or Y-cut-off.[6] Suppose that non-constant slope is allowed for `X1` and `X2`. The S-plus notation for the extended model would be

$$\texttt{y} \sim \texttt{cohort} * (\texttt{X1} + \texttt{X2}) + \texttt{X3}$$

The interaction notation ($*$) implies that lower-order effects are also included in the model. The extended CR model is a discrete version of the Cox survival model with time-dependent covariables.

The `cr.setup` function in `Design` returns a list of vectors useful in constructing a data set used to 'trick' a binary logistic function into fitting CR models. The `subs` vector in this list contains observation numbers in the original data, some of which are repeated so that subscripting on `subs` will cause the subject's row of predictor variable values to be replicated the desired number of times ($\min(Y + 1, k)$ times if $Y = 0, 1, \ldots, k$). Each replication corresponds to the conditioning event or risk set (cohort) $Y \geqslant j$, and the new dummy response variable `y` indicates whether or not $Y = j$.

```
u ← cr.setup(Y)          # Y is original ordinal response vector
attach(mydata[u$subs,]) # my data is the original dataset
                         # mydata[i,] subscripts the input data,
                         # using duplicate values of i for repeats
y      ← u$y             # constructed binary response
cohort ← u$cohort        # cohort or risk set categories
```

Here the cohort variable has values 'all', 'Y > = 1' corresponding to the conditioning events in equation (10). After the attach command runs, vectors such as age are lengthened (to 5553 records) by duplicating the correct observations according to the magnitude of a subject's Y value. Now we fit a fully extended CR model which makes no equal slopes assumptions, that is, the model *has* to fit Y assuming the covariables are linear and additive. At this point, we omit hrat but add back all variables which were deleted by examining their association with Y. Recall that most of these 7 cluster scores were summarized using PC_1. Adding back 'insignificant' variables will allow us to validate the model fairly using the bootstrap, as well as to obtain confidence intervals which are not falsely narrow.[50]

```
full ← lrm(y ~ cohort*(ageg*(rcs(temp,5) + rcs(rr,5)) + rcs(waz,4) +
                       bul.conv + drowsy + agitated + reffort +
                       ausc + feeding + abdominal +
                       hydration + hxprob + pustular + crying +
                       fever.ill + stop.breath + labor), x=T, y=T)
# x=T, y=T is for pentrace, validate, calibrate below
latex(anova(full))
```

This model has LR $\chi^2 = 1824$ with 87 d.f. Wald statistics produced by anova(full) are in Table VI. For brevity, tests of non-linear effects and many tests with $P > 0{\cdot}1$ are not shown.

The global test of the constant slopes assumption in the CR model (test of all interactions involving cohort) has $\chi^2 = 172$ with 43 d.f., $P < 0{\cdot}0001$. Consistent with Figure 5, the formal tests indicate that ausc is the biggest violator, followed by waz and rr.

At this point we select the CR model for this problem because of its flexibility, both in testing the equal slopes assumption and in parameterizing non-equal-slopes extensions.

11. PENALIZED ESTIMATION

The traditional estimation technique used for logistic models, maximum likelihood estimation (MLE), is optimal (that is, has lowest variance) among techniques which yield unbiased estimates for large samples. The bias of an estimator is not its most important attribute, however. The probability that a parameter estimate $\hat{\beta}_i$ is close to the true population value is a function of the mean squared error of that estimate, which is the variance plus the square of the bias. In many cases, especially for small samples, one can sacrifice the bias and lower the variance by a sufficient amount so that the mean squared error of the estimate is lower than that of the MLE. As an example, if one were using patients' sex to predict mortality and there were two males in the sample, the probability of death for males would be more reliably estimated by 'shrinking' it toward the probability of death for the overall sample, which is dominated by females. Shrinkage can be much more general, for example, shrinking a non-linear regression effect toward a linear effect if the evidence for non-linearity is weak.

Table VI. Wald statistics for y in the extended CR model

	χ^2	d.f.	P
cohort	199·47	44	< 0·0001
All interactions	172·12	43	< 0·0001
ageg	48·89	36	0·0742
temp	59·37	24	0·0001
rr	93·77	24	< 0·0001
waz	39·69	6	< 0·0001
bul.conv	10·80	2	0·0045
drowsy	15·19	2	0·0005
agitated	13·55	2	0·0011
reffort	51·85	2	< 0·0001
ausc	109·80	2	< 0·0001
feeding	27·47	2	< 0·0001
hxprob	6·62	2	0·0364
stop.breath	5.34	2	0·0693
labor	5·35	2	0·0690
ageg × temp	8·18	16	0·9432
ageg × rr	38·11	16	0·0015
cohort × rr	19·67	12	0·0736
cohort × waz	9·04	3	0·0288
cohort × ausc	38·11	1	< 0·0001
cohort × fever.ill	3·17	1	0·0749
cohort × stop.breath	2·99	1	0·0839
cohort × ageg × temp	2·22	8	0·9736
cohort × ageg × rr	10·22	8	0·2500
Total non-linear	93·36	40	< 0·0001
Total interaction	203·10	59	< 0·0001
Total non-linear + interaction	257·70	67	< 0·0001
Total	1211·73	87	< 0·0001

Penalized MLE (PMLE)[51–53] is a general technique for shrinking (stabilizing) regression fits. Instead of maximizing the log-likelihood, PMLE maximizes a penalized log-likelihood which is the sum of the ordinal model log-likelihood and a penalty, resulting in

$$\log L - \frac{1}{2}\lambda \sum_{i=1}^{p} (s_i\beta_i)^2. \tag{12}$$

Here $s_1, s_2, \ldots, s_p$ are scale factors chosen to make $s_i\beta_i$ unitless. Most authors standardize the data first and do not have scale factors in the equation,[51] but equation (12) has the advantage of allowing estimation of β on the original scale of the data. The usual methods (for example, Newton–Raphson) are used to maximize equation (12). The usual default values for s are sample standard deviations of columns of the design matrix, but special consideration has to be given to dummy variables,[52] which gives rise to a more general form of the penalized log-likelihood

$$\log L - \frac{1}{2}\lambda\beta' P\beta \tag{13}$$

where P is a penalty matrix. Rows and columns of P can easily be set to zero for parameters for which no shrinkage is desired.[52,53]

The main problem in using PMLE is the choice of λ. Many authors use cross-validation to solve for the λ which optimizes an unbiased estimate of predictive accuracy, but it is easy to show that one must use a huge number of data splits to get a precise estimate of the optimum λ. A faster and usually more reliable strategy, based on findings from a small number of simulation studies, is to choose the λ which maximizes the 'effective' AIC. Gray (Eq. 2·9)[53] and others show how to compute the 'effective d.f.' in this situation (that is, higher λ causes more shrinkage which lowers the effective d.f.). The effective AIC is

$$\text{LR } \chi^2 - 2 \times \text{effective d.f.} \tag{14}$$

where LR χ^2 is the likelihood ratio χ^2 for the penalized model, but ignoring the penalty function.

The lrm function will do PMLE, and a separate function called pentrace searches for the optimum λ based on effective AIC once the analyst specifies a vector of λs to try. pentrace can also allow for differing λ for different types of terms in the model. Here we want to do a grid search to determine the optimum penalty for simple main effect (non-interaction) terms and the penalty for interaction terms, most of which are terms interacting with cohort to allow for unequal slopes. The following code uses pentrace on the full extended CR model fit to find the optimum penalty factors. All combinations of simple and interaction λ's for which the interaction penalty $\geqslant$ the penalty for the simple parameters are examined. The range of penalty factors to try for each type of parameter was found by computing effective AIC in a trial and error process.

```
pentrace(full, list(simple=c(0,.025,.05,.075,.1),
                    interaction=c(0,10,50,100,125,150)))
Best penalty:

  simple interaction     df     aic
    0.05         125  49.75 1672.6

  simple  interaction     df     aic     bic   aic.c
   0.000            0 87.000 1650.3 1074.2 1647.5
   0.000           10 60.628 1670.8 1269.4 1669.5
   0.025           10 60.110 1671.6 1273.5 1670.2
   0.050           10 59.797 1671.6 1275.6 1670.3
   0.075           10 59.581 1671.5 1276.9 1670.2
   0.100           10 59.421 1671.3 1277.8 1670.0
   0.000           50 54.640 1671.3 1309.5 1670.2
   0.025           50 54.135 1672.0 1313.5 1670.9
   0.050           50 53.829 1672.0 1315.5 1670.9
   0.075           50 53.619 1671.8 1316.8 1670.8
   0.100           50 53.463 1671.6 1317.6 1670.5
   0.000          100 51.613 1671.9 1330.1 1670.9
   0.025          100 51.113 1672.5 1334.1 1671.6
   0.050          100 50.809 1672.6 1336.1 1671.6
   0.075          100 50.600 1672.4 1337.3 1671.4
   0.100          100 50.445 1672.1 1338.1 1671.2
   0.000          125 50.553 1671.9 1337.2 1671.0
```

```
0.025      125 50.054 1672.6 1341.1 1671.7
0.050      125 49.750 1672.6 1343.2 1671.7
0.075      125 49.542 1672.4 1344.4 1671.5
0.100      125 49.387 1672.1 1345.1 1671.3
0.000      150 49.653 1671.8 1343.0 1670.9
0.025      150 49.155 1672.5 1347.0 1671.6
0.050      150 48.852 1672.5 1349.0 1671.6
0.075      150 48.643 1672.3 1350.2 1671.5
0.100      150 48.489 1672.1 1351.0 1671.2
```

We see that shrinkage from 87 d.f. down to 49·8 effective d.f. results in an increase in AIC of 22·3. The optimum penalty factors were 0·05 for simple terms and 125 for interaction terms.*

We now store a penalized version of the full fit, determine the kind of model terms for which the effective d.f. were reduced, and compute χ^2 for each factor in the model. We take the effective d.f. for a collection of model parameters to be the sum of the diagonals of the matrix product defined underneath Gray's Eq. 2·9[53] that correspond to those parameters:

```
full.pen ← update(full, penalty=list(simple=.05, interaction=125))
effective.df(full.pen)
Original and Effective Degrees of Freedom
                              Original    Penalized
                   All              87        49.75
         Simple Terms               20        19.98
Interaction or Nonlinear            67        29.77
             Nonlinear              40        16.82
           Interaction              59        22.57
Nonlinear Interaction               32         9.62
plot(anova(full.pen))
somers2(predict(full.pen)[cohort=='all'], y[cohort=='all'])
     C    Dxy       n Missing
 0.836  0.672    4554       0
```

This will be the final model except for the model used in Section 12. The model has LR $\chi^2 = 1772$. The output of effective.df shows that non-interaction terms have barely been penalized, and coefficients of interaction terms have been shrunken from 59 d.f. to effectively 22·6 d.f. Predictive discrimination was assessed by computing the Somers' D_{xy} rank correlation between $X\hat{\beta}$ and whether or not $Y = 0$, in the subset of records for which $Y = 0$ is what was being predicted. Here $D_{xy} = 0{\cdot}672$ and the ROC area is 0·836 (the unpenalized model had an apparent $D_{xy} = 0{\cdot}676$ for the training sample). To summarize in another way, the effectiveness of this model in screening

* See Hurvich and Tsai[54] for the definition of aic.c, the 'corrected AIC'. Regarding the Bayesian information criterion (bic) of Schwarz,[55] several simulations have shown that models selected by BIC had too much shrinkage and hence in validation samples predicted less well than ones selected using AIC

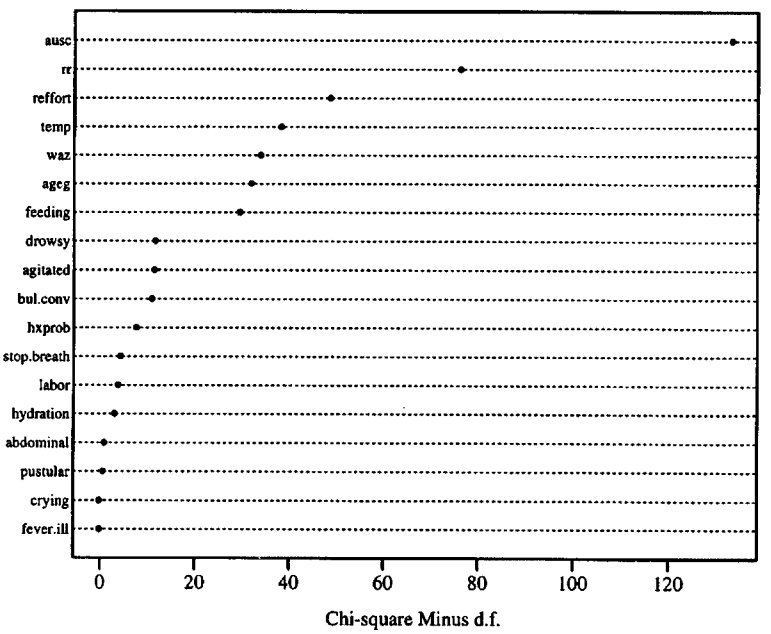


Figure 6. Importance of predictors in full penalized model, as judged by partial Wald χ^2 minus the predictor d.f. The Wald χ^2 values for each line in the dot plot include contributions from all higher-order effects. For example, the cohort effect includes all cohort interaction effects

infants for risks of any abnormality, the fraction of infants with predicted probabilities that $Y > 0$ being $< 0{\cdot}05$, $> 0{\cdot}25$, and $> 0{\cdot}5$ are, respectively, 0·10, 0·28 and 0·14. anova output is plotted in Figure 6 to give a snapshot of the importance of the various predictors. The Wald statistics used here are computed on a variance-covariance matrix which is adjusted for penalization (Gray Eq. 2·6[53]).

The full equation for the fitted model is obtained using the S-plus statement `latex(full.pen)`, which typeset the output below using LaTeX.* Restricted cubic spline functions which were fit using rcs have been automatically written in more simple unrestricted form (Herndon and Harrell, Eq. 1[41]). Only the part of the equation used for predicting $\Pr(Y = 0)$ is shown:

$$\Pr\{Y = 0\} = \frac{1}{1 + \exp(-X\beta)}$$

where

$$X\hat{\beta} = -5{\cdot}543 + 0{\cdot}1075\{\text{ageg} \in [7,60)\} + 0{\cdot}1971\{\text{ageg} \in [60,90]\}$$
$$+ 0{\cdot}1979\,\text{temp} + 0{\cdot}1092(\text{temp} - 36{\cdot}2)_+^3 - 2{\cdot}833(\text{temp} - 37)_+^3 + 5{\cdot}071(\text{temp} - 37{\cdot}3)_+^3$$

* The fitted equation could be written in the S-plus or SAS language using Design's Function function had third-order interactions not been present

$$
\begin{aligned}
&- 2{\cdot}508(\text{temp} - 37{\cdot}7)_+^3 + 0{\cdot}1606(\text{temp} - 39)_+^3 \\
&+ 0{\cdot}02091\,\text{rr} - 6{\cdot}337 \times 10^{-5}(\text{rr} - 32)_+^3 + 8{\cdot}405 \times 10^{-5}(\text{rr} - 42)_+^3 + 6{\cdot}152 \times 10^{-5}(\text{rr} - 49)_+^3 \\
&- 0{\cdot}0001018(\text{rr} - 59)_+^3 + 1{\cdot}96 \times 10^{-5}(\text{rr} - 76)_+^3 \\
&- 0{\cdot}0759\,\text{waz} + 0{\cdot}02509(\text{waz} + 2{\cdot}9)_+^3 - 0{\cdot}1185(\text{waz} + 0{\cdot}75)_+^3 + 0{\cdot}1226(\text{waz} - 0{\cdot}28)_+^3 \\
&- 0{\cdot}02916(\text{waz} - 1{\cdot}73)_+^3 - 0{\cdot}4418\,\text{bul.conv} - 0{\cdot}08185\,\text{drowsy} - 0{\cdot}05327\,\text{agitated} \\
&- 0{\cdot}2304\,\text{reffort} - 1{\cdot}159\,\text{ausc} - 0{\cdot}16\,\text{feeding} - 0{\cdot}1609\,\text{abdominal} \\
&- 0{\cdot}0541\,\text{hydration} + 0{\cdot}08086\,\text{hxprob} + 0{\cdot}00752\,\text{pustular} + 0{\cdot}04712\,\text{crying} \\
&+ 0{\cdot}004299\,\text{fever.ill} - 0{\cdot}3519\,\text{stop.breath} + 0{\cdot}06864\,\text{labor} \\
&+ \{\text{ageg} \in [7,60)\}[6{\cdot}5 \times 10^{-5}\,\text{temp} - 0{\cdot}0028(\text{temp} - 36{\cdot}2)_+^3 - 0{\cdot}008691(\text{temp} - 37)_+^3 \\
&- 0{\cdot}004988(\text{temp} - 37{\cdot}3)_+^3 + 0{\cdot}02592(\text{temp} - 37{\cdot}7)_+^3 - 0{\cdot}009445(\text{temp} - 39)_+^3] \\
&+ \{\text{ageg} \in [60,90]\}[0{\cdot}000132\,\text{temp} - 0{\cdot}001826(\text{temp} - 36{\cdot}2)_+^3 - 0{\cdot}0164(\text{temp} - 37)_+^3 \\
&- 0{\cdot}0476(\text{temp} - 37{\cdot}3)_+^3 + 0{\cdot}09142(\text{temp} - 37{\cdot}7)_+^3 - 0{\cdot}02559(\text{temp} - 39)_+^3] \\
&+ \{\text{ageg} \in [7,60)\}[-0{\cdot}0009438\,\text{rr} - 1{\cdot}045 \times 10^{-6}(\text{rr} - 32)_+^3 - 1{\cdot}67 \times 10^{-6}(\text{rr} - 42)_+^3 \\
&- 5{\cdot}189 \times 10^{-6}[\text{rr} - 49)_+^3 + 1{\cdot}429 \times 10^{-5}(\text{rr} - 59)_+^3 - 6{\cdot}382 \times 10^{-6}(\text{rr} - 76)_+^3] \\
&+ \{\text{ageg} \in [60,90]\}[-0{\cdot}001921\,\text{rr} - 5{\cdot}521 \times 10^{-6}(\text{rr} - 32)_+^3 - 8{\cdot}628 \times 10^{-6}(\text{rr} - 42)_+^3 \\
&- 4{\cdot}147 \times 10^{-6}(\text{rr} - 49)_+^3 + 3{\cdot}813 \times 10^{-5}(\text{rr} - 59)_+^3 - 1{\cdot}984 \times 10^{-5}(\text{rr} - 76)_+^3]
\end{aligned}
$$

and $\{c\} = 1$ if subject is in group c, otherwise, $(x)_+ = x$ if $x > 0$, 0 otherwise.

To show the shapes of effects of the predictors we use the following code. For the continuous variables temp and rr which interact with age group, we show the effects for all three age groups, separately for each Y cut-off. All effects have been centred so that the log odds at the median predictor value is zero when cohort='all', so these plots actually show log odds relative to the reference values. The patterns in Figure 7 are in agreement with those in Figure 5:

```
par(mfrow=c(4,4))
yl←c(−2.5, 1)  # put all plots on common y-axis scale

# Plot predictors which interact with another predictor
# Vary ageg over all age groups, then vary temp over its
# default range (10th smallest to 10th largest values in data)
# Make a separate plot for each 'cohort'
# ref.zero centers effects using median x

for(co in levels(cohort)) {
  plot(full.pen, temp=NA, ageg=NA, cohort=co, ref.zero=T, ylim=yl, conf.int=F)
  text(37.5, 1.5, co)        # add title showing current cohort
}
```

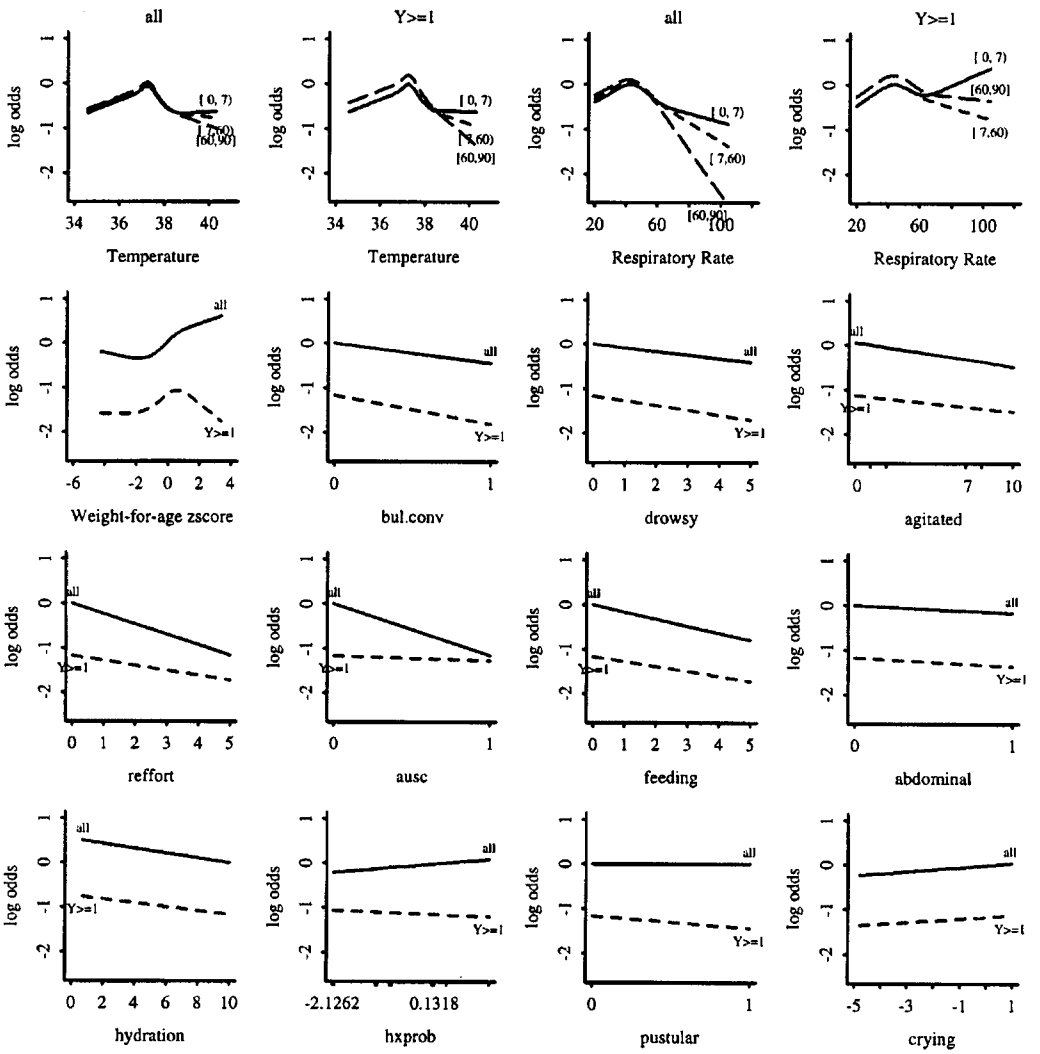


Figure 7. Centred effects of predictors on the log odds. The first four plots show interaction effects with the age intervals noted. For others, interaction with cohort are shown. For predictors having fewer than 10 unique values, *x*-axis tick marks appear only for values which occurred in the data. No plot was made for the fever.ill, stop.breath, or labor cluster scores. The title all refers to the prediction of $Y = 0 | Y \geq 0$, that is, $Y = 0$

```
for(co in levels(cohort)) {
  plot(full.pen, rr=NA, ageg=NA, cohort=co, ref.zero=T, ylim=yl, conf.int=F)
  text(70, 1.5, co)
}

# For each predictor which only interacts with cohort, show the
# Differing effects of the predictor for predicting Pr(Y=0) and
```

```
# Pr(Y=1 | Y > 0) on the same graph

plot(full.pen, waz       = NA, cohort=NA, ref.zero=T, ylim=yl, conf.int=F)
plot(full.pen, bul.conv = NA, cohort=NA, ref.zero=T, ylim=yl, conf.int=F)
. . . .
plot(full.pen, crying    = NA, cohort=NA, ref.zero=T, ylim=yl, conf.int=F)
```

12. USING APPROXIMATIONS TO SIMPLIFY THE MODEL

It is tempting to use P-values and stepwise methods to develop a parsimonious prediction model. Besides invalidating confidence limits and causing measures of predictive accuracy such as adjusted R^2 to be optimistic, there are many other reasons not to rely on stepwise techniques (see Harrell *et al.*[20] for citations). We follow Spiegelhalter's advice to use full model fits in conjunction with shrinkage.[56]

Parsimonious models can be developed, however, by approximating predictions from the model to any desired level of accuracy. Let $\hat{L} = X\beta$ denote the predicted log odds from the full penalized ordinal model, including multiple records for subjects with $Y > 0$. Then we can use a variety of techniques to approximate $\hat{L}$ from a subset of the predictors (in their raw form). With this approach one can immediately see what is lost over the full model by computing, for example, the mean absolute error in predicting $\hat{L}$. Another advantage to full model approximations is that shrinkage used in computing $\hat{L}$ is inherited by any model that predicts $\hat{L}$. In contrast, the usual stepwise methods result in $\hat{\beta}$ that are too large since the final coefficients are estimated as if the model structure was prespecified.*

Even though CART (classification and regression trees[58]) when used on X and Y often finds prediction rules that validate poorly because of the extremely large number of models searched,[27] CART can be very useful as an approximator for a complex model. For the current problem, CART would be particularly useful as it would result in a prediction tree that would be easy for health workers to use. Unfortunately, a 50-node CART was required to predict $\hat{L}$ with an $R^2 \geqslant 0{\cdot}9$, and the mean absolute error in the predicted logit was still 0·4. This will happen when the model contains many important continuous variables.

We chose to approximate the full model using its important components, by using a stepdown technique predicting $\hat{L}$ from all of the component variables using ordinary least squares. In using stepdown with the least squares function ols in Design there is a problem with infinite F statistics when the initial $R^2 = 1{\cdot}0$, so we will specify $\sigma = 1$ to ols. Because cohort interacts with the predictors, separate approximations can be developed for each level of Y. For this example we approximate the log odds that $Y = 0$ using the cohort of patients used for determining $Y = 0$, that is, $Y \geqslant 0$ or cohort='all':

```
plogit ← predict(full.pen)

f ← ols(plogit ~ ageg*(rcs(temp,5) + rcs(rr,5)) + rcs(waz,4) +
        bul.conv + drowsy + agitated + reffort + ausc + feeding +
```

* The *lasso* method of Tibshirani[57] addresses this problem

```
        abdominal + hydration + hxprob + pustular + crying +
        fever.ill + stop.breath + labor, sigma=1,
        subset=cohort='all')

# Do fast backward stepdown
fastbw(f, aics=1e10)   # 1e10 causes all variables to eventually be
                       # deleted so can see most important ones in order

# Fit an approximation to the full penalized model using most
# important variables
full.approx ← ols(plogit ~ rcs(temp,5) + ageg*rcs(rr,5) + rcs(waz,4) +
                  bul.conv + drowsy + reffort + ausc + feeding,
                  subset=cohort=='all')
```

The approximate model had R^2 against the full penalized model of 0·972, and the mean absolute error in predicting $\hat{L}$ was 0·17. The D_{xy} rank correlation between the approximate model's predicted logit and the binary event $Y = 0$ is 0·665 as compared with the full model's $D_{xy} = 0{\cdot}672$.

Next, turn to diagramming this model approximation so that all predicted values can be computed without the use of a computer. We draw a type of nomogram which converts each effect in the model to a 0–100 scale which is just proportional to the log odds. These points are added across predictors to derive the 'total points', which are converted to $\hat{L}$ and then to predicted probabilities. For the interaction between rr and ageg, Design's nomogram function automatically constructs 3 rr axes – only one is added into the total point score for a given subject. Here we draw a nomogram for predicting the probability that $Y > 0$, which is $1 - \Pr(Y = 0)$. This probability is derived by negating $\hat{\beta}$ and $X\hat{\beta}$ in the model derived to predict $\Pr(Y = 0)$.

```
f ← full.approx
f$coefficients       ← f$coefficients
f$linear-predictors ← f$linear.predictors

nomogram(f,
         temp=32:41, rr=seq(20,120,by=10), waz=seq(-1.5,2,by=.5),
         fun=plogis, funlabel='Pr(Y > 0)',
         fun.at=c(.02,.05,seq(.1,.9,by=.1),.95,.98))
# plogis is S-PLUS's builtin 1/(1+exp(-x)) function
```

The nomogram is shown in Figure 8. As an example in using the nomogram, a 6-day old infant gets approximately 9 points for having a respiratory rate of 30/min, 19 points for having a temperature of 39°C, 11 points for waz=0, 14 points for drowsy=5, and 15 points for reffort=2. Assuming that bull.conv=ausc=feeding=0, that infant gets 68 total points. This corresponds to $X\hat{\beta} = -0{\cdot}6$ and a probability of 0·35. Values computed directly from the full.approx formula were $X\hat{\beta} = -0{\cdot}68$ and a probability of 0·34.*

* To see how this compares with prediction using the full model, the extra clinical signs in that model that are not in the approximate model were predicted individually on the basis of $X\hat{\beta}$ from the reduced model along with the signs that are in that model, using ordinary linear regression. The signs not specified when evaluating the approximate model were then set to predicted values based on the values given for the 6-day old infant above. The resulting $X\hat{\beta}$ for the full model is $-0{\cdot}81$ and the predicted probability is 0·31, as compared with $-0{\cdot}68$ and 0·34 quoted above

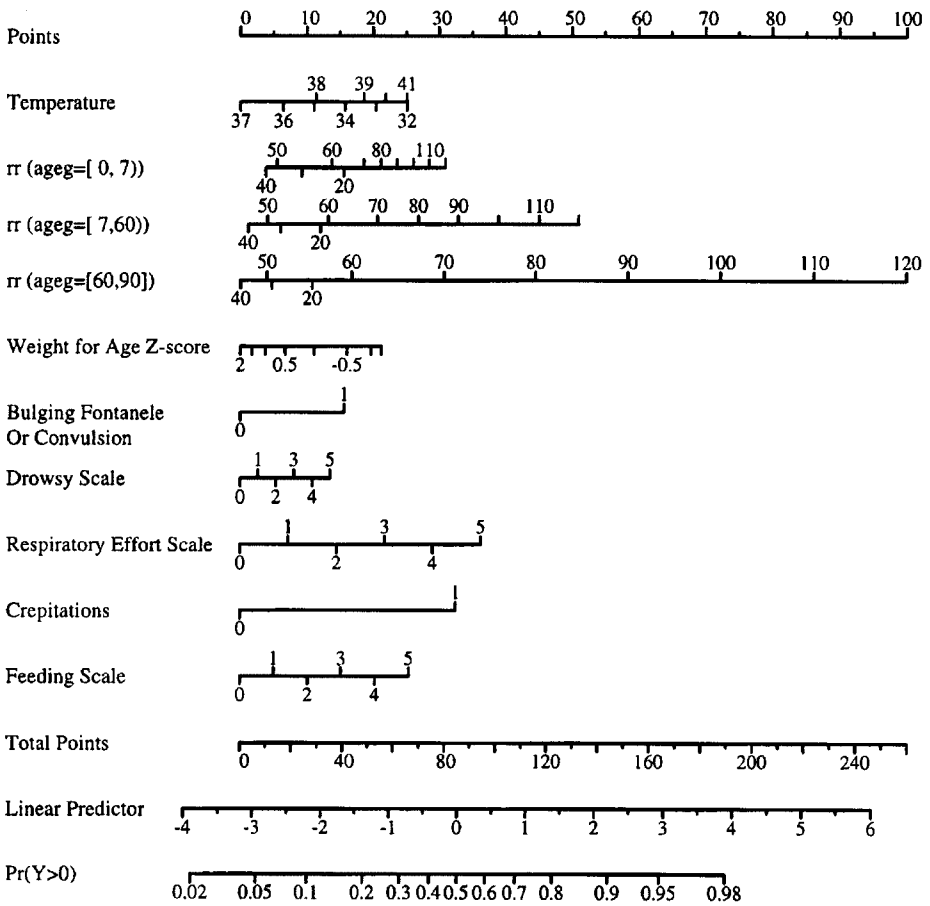


Figure 8. Nomogram for predicting $\Pr(Y > 0)$ from the penalized extended CR model, using an approximate model fitted using ordinary least squares ($R^2 = 0{\cdot}972$ against the full model's predicted logits)

For some applications a further simplification of the model is required. Assuming that maximum information is to be derived from the important continuous variables, one way to simplify the model is to dichotomize each clinical sign into a present/absent coding. To find the simplest approximation of the model that adequately discriminated between low- and high-risk infants, we again predicted the 'good standard' predicted log odds of outcomes, but used as candidate variables the 3-interval age variable, the vital signs, vital sign by age interactions, and all of the individual clinical signs and clinical history variables. To enable the procedure to automatically find the best cutpoints for multi-level signs, those signs were represented using a series of binary variables. For example, `hfa` (history of feeding) is a 0–4 variable, and the following candidate variables were used to represent `hfa`: mildly reduced or worse; severely reduced or unable to feed, and unable to feed. All variables were fitted in an ordinary multiple regression model, and variables were deleted in increasing order of explained variation. A model

which retained 7 individual signs resulting in $D_{xy} = 0{\cdot}664$ and R^2 against the optimal model's predicted logit of 0·954.

13. VALIDATING THE MODEL

Most analysts validate a fitted model using held-back data, but this method has severe drawbacks.[20] The bootstrap technique[59] allows the analyst to derive bias (overfitting) – corrected estimates of predictive accuracy without holding back valuable data during the model development phase. The steps required for using the bootstrap to bias-correct indexes such as D_{xy} and calibration error was summarized in Harrell *et al.*[20] For the full CR model which was fitted using PMLE, we used 150 bootstrap replications to estimate and then to correct for optimism in various statistical indexes: D_{xy}; generalized R^2;[60] intercept and slope of a linear recalibration equation for $X\hat{\beta}$ (related to Section 7 of van Houwelingen and le Cessie;[61] see also Phillips *et al.*[62]); the maximum calibration error for $\Pr(Y = 0)$ based on the linear-logistic recalibration (Emax), and the Brier quadratic probability score B.[63] PMLE is used at each of the 150 resamples. During the bootstrap simulations, we sample with replacement from the *patients* and not from the 5553 expanded *records*, hence the specification cluster=u$subs, where u$subs is the vector of sequential patient numbers computed from cr.setup above. To be able to measure the predictive accuracy of the predicted probability of a single event, the subset parameter is specified so that $\Pr(Y = 0)$ is being assessed even though 5553 observations are used to develop each of the 150 models. The output and the S-plus statement used to obtain the output are shown below:

```
validate(full.pen, B=150, cluster=u$subs, subset=cohort=='all')
```

	index.orig	training	test	optimism	index.corrected	n
Dxy	0.672	0.675	0.666	0.009	0.662	150
R2	0.376	0.383	0.370	0.013	0.363	150
Intercept	−0.031	−0.033	0.001	−0.034	0.003	150
Slope	1.029	1.031	1.002	0.029	1.000	150
Emax	0.000	0.000	0.001	0.001	0.001	150
B	0.120	0.119	0.121	−0.002	0.122	150

We see that for the apparent $D_{xy} = 0{\cdot}672$ the optimism from overfitting was estimated to be 0·009 for the PMLE model, so the bias-corrected estimate of predictive discrimination is 0·662. The intercept and slope needed to recalibrate $X\hat{\beta}$ to a 45° line are very near (0, 1). The estimate of the maximum calibration error in predicting $\Pr(Y = 0)$ is 0·001 which is quite satisfactory. The corrected Brier score is 0·122.

The simple calibration statistics just listed do not address the issue of whether predicted values from the model are miscalibrated in a non-linear way. Steps for estimating bias-corrected calibration curves for survival time models and for non-parametrically estimating a smooth calibration curve for a binary logistic model on a separate validation sample were given previously.[20] Putting these two techniques together we arrive at the following plan for estimating a calibration curve using the bootstrap, with the only assumption being the smoothness of the curve. Choose a single binary event for which to check the calibration of the estimated

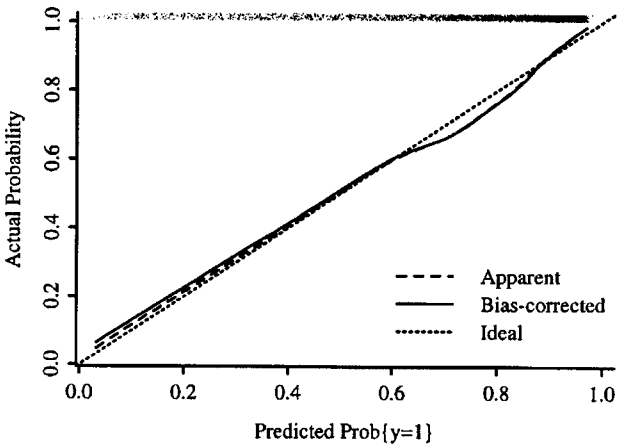


Figure 9. Bootstrap calibration curve for the full penalized extended CR model. 150 bootstrap repetitions were used in conjunction with the lowess smoother.[49] Also shown is a 'rug plot' to demonstrate how effetive this model is in discriminating patients into low and high risk groups for $\Pr(Y = 0)$ (which corresponds with the derived variable value $y = 1$ when cohort = 'all')

probabilities. The actual occurrences of binary responses are smoothed using lowess (with the 'no iteration' option) to estimate probabilities. Then choose a grid of predicted values, for example, 0·01, 0·03, 0·05, ... , 0·99. Fit lowess to the predicted probabilities derived from the final model and the actual binary outcomes from the final model and the actual binary outcomes from the original sample. Then evaluate the smoothed estimates at the grid. Differences between the lowess estimates and the 45° line are the estimates of apparent calibration accuracy. Then for each bootstrap resample, the ordinal model is fitted using PMLE from a sample with replacement from the *patients*, and the coefficients from this model are used to predict probabilities for the original sample. The discrepancies from the 45° line are compared with the discrepancies present when the bootstrap model was evaluated on the bootstrap sample. The difference in the discrepancies is the estimate of optimism. After averaging over 150 replications, separately for each probability level in the uniform grid, the estimates of optimism in the original, apparent, calibration errors are added to those errors. Then the bootstrap-corrected calibration curve is plotted.

All these steps are done using the following Design functions:

```
cal ← calibrate(full.pen, B = 150, cluster = u$subs, subset = cohort = = 'all')
plot(cal)
```

The results are shown in Figure 9. One can see a slightly non-linear calibration function estimate, but the overfitting-corrected calibration is excellent everywhere, being only slightly worse than the apparent calibration. The estimated maximum calibration error is 0·043. The excellent validation for both predictive discrimination and calibration are a result of the large sample size, frequency distribution of Y, initial data reduction and PMLE.

14. SUMMARY

Clinically-guided variable clustering and item weighting, done with very limited use of the outcome variable, resulted in a great reduction in the number of candidate predictor degrees of freedom and hence increased the true predictive accuracy of the model. Sources summarizing clusters of clinical signs, along with the temperature, respiratory rate, and weight-for-age after suitable non-linear transformation and allowance for interactions with age, are powerful predictors of the ordinal response. Graphical methods are effective for detecting lack of fit in the PO and CR models and for diagramming the final model. Model approximation is a better approach than stepwise methods (that use Y) to develop parsimonious clinical prediction tools. Approximate models inherit the shrinkage from the full model. For the ordinal model developed here, substantial shrinkage of the full model was needed.

The bootstrap, as in a wide variety of other situations, is an effective tool for validating an ordinal logistic model with respect to discrimination and calibration without having the need to hold back data during model development. The final CR ordinal logistic model accurately predicted severity of diagnosis/outcome (as summarized by several disparate outcome variables) in infants screened for pneumonia, sepsis, and meningitis in developing countries. There was nothihg about the continuation ratio model that made it fit the data set better than other ordinal models (which we have found to be the case in one other large data set), and in fact there is some evidence that the equal-slopes CR model fits the data more poorly than the equal-slopes PO model. The real benefit of the CR model is that using standard binary logistic model software one can flexibly specify how the equal-slopes assumption can be relaxed.

Faraway[64] has demonstrated how all data-driven steps of the modelling process increase the real variance in 'final' parameter estimates, when one estimates variances without assuming that the final model was prespecified. For ordinal regression modelling, the most important modelling steps are (i) choice of predictor variables; (ii) selecting or modelling predictor transformations; and (iii) allowance for unequal slopes across Y-cut-offs (that is, non-PO or non-CR). Regarding steps (ii) and (iii) one is tempted to rely on graphical methods such as residual plots to make detours in the strategy, but it is very difficult to estimate variances or to properly penalize assessments of predictive accuracy for subjective modelling decisions. Regarding (i), shrinkage has been proven to work better than stepwise variable selection when one is attempting to build a main-effects model.[56] Choosing a shrinkage factor is a well-defined, smooth, and often a unique process as opposed to binary decisions on whether variables are 'in' or 'out' of the model. Likewise, instead of using arbitrary subjective (residual plots) or objective (χ^2 due to cohort × covariable interactions, that is, non-constant covariable effects) assessments, shrinkage can systematically allow model enhancements in so far as the information content in the data will support, through the use of differential penalization. Shrinkage is a solution to the dilemma faced when the analyst attempts to choose between a parsimonious model and a more complex one that fits the data. Penalization does not require the analyst to make a binary decision, and it is a process that can be validated using the bootstrap.

APPENDIX

WHO/ARI Young Infant Multicentre Study Group*

Study sites	Addis Ababa, Ethiopia	Fajara, Gambia	Goroka, Papua New Guinea	Manila, Philippines
Principal investigator	Lulu Muhe	Kim Mulholland	Deborah Lehmann	Salvacion Gatchalian
Co-investigators, Study coordinators		Olayinka Ogunlesi Martin Weber	Gerard Saleu	Beatriz Quiambao
Other investigators, clinicians	Meaza Tilahun Sileshi Lulseged Senait Kebede	Mark Manary Ayo Palmer	Alphonse Rongap Mexy Kakazo Pioto Namuigi Sebeya Lupiwa Rebecca sehuko	Ana Marie Moreles Leticia Abraham
Bacteriologists	Afeworti Yohanes† Bahrie Belete Signe Ringertz	Richard Adegbola Osman Secka	Alison Clegg Audrey Michael Tony Lupiwa Matthew Omena Mark Mens	Lydia Sombrero Ma. Victoria Abraham
Radiologists	Tsegaye Desta			
Data management	Kidanemariam Wlyesus	Joseph Bangali	Don Lewis	Elinor S. Sunico Teresita C. Cedulla
Institution/ hospitals	Department of Paediatrics and Child Heath, Ethio-Swedish Children's Hospital, Addis Ababa University	Medical Research Council Hospital and Royal Victoria Hospital	Papua New Guinea Institute of Medical Research Goroka Base Hospital	Research Institute of Tropical Tropical Medicine, Phillippines General Hospital, Quezon City General Hospital
Institution Directors	Nebiat Tafari	Brian Greenwood	Michael P. Alpers	

* Not inclusive
† Deceased

Study Co-ordination

Scientific co-ordinator: Dr Sandy Gove, WHO/ARI
Data management: Dr Peter Byass, University of Nottingham Medical School
Data analysis: Dr Frank Harrell, University of Virginia, Charlottesville Mrs Karen Mason, WHO/ARI
Management, logistic, supplies: Mrs Frances McCaul, WHO/ARI
Mrs Sue Parker, WHO/ARI
Study advisors: Dr Claire Broome, Centers for Disease Control (CDC), Atlanta
Dr H. F. Eichenwald, University of Texas Southwestern Medical Center, Dallas
Mr Mike Gratten, Queensland Institute of Medical Research, Brisbane
Dr P. Margolis, University of North Carolina at Chapel Hill
Dr R. Facklam, CDC, Atlanta
Radiology working group: Dr H. Tschappeler, Universitat Bern
Dr A. Lamont, The Leicester Royal Infirmary
Dr G. M. A. Hendry, Royal Hospital for Sick Children, Edinburgh
Professor Philip E. S. Palmer, University of California, Davis

ACKNOWLEDGEMENTS

This work was supported by The World Health Organization ARI Programme, and for F. Harrell, Research Grants HS-06830 and HS-07137 from the Agency for Health Care Policy and Research, Rockville, Maryland, U.S.A., and grants from the Robert Wood Johnson Foundation,

Princeton, NJ, U.S.A. F. Harrell wishes to dedicate his work on this project to the memory of his dear colleague L. Richard Smith whose critical reading of this paper resulted in significant improvements.

REFERENCES

1. Walker, S. H. and Duncan, D. B. 'Estimation of the probability of an event as a function of several independent variables', *Biometrika*, **54**, 167–178 (1967).
2. Fienberg, S. E. *The Analysis of Cross-Classified Data*, 2nd edn, MIT Press, Cambridge, MA, 1980.
3. Agresti, A. 'A survey of models for repeated ordered categorical response data', *Statistics in Medicine*, **8**, 1209–1224 (1989).
4. Anderson, J. A. and Philips, P. R. 'Regression, discrimination and measurement models for ordered categorical variables', *Applied Statistics*, **30**, 22–31 (1981).
5. Anderson, J. A. 'Regression and ordered categorical variables', *Journal of the Royal Statistical Society*, Series B, **46**, 1–30 (1984).
6. Armstrong, B. G. and Sloan, M. 'Ordinal regression models for epidemiologic data', *American Journal of Epidemiology*, **129**, 191–204 (1989).
7. Ashby, D., West, C. R. and Ames, D. 'The ordered logistic regression model in psychiatry: Rising prevalence of dementia in old people's homes', *Statistics in Medicine*, **8**, 1317–1326 (1989).
8. Berridge, D. M. and Whitehead, J. 'Analysis of failure time data with ordinal categories of response', *Statistics in Medicine*, **10**, 1703–1710 (1991).
9. Brazer, S. R., Pancotto, F. S., Long III, T. T., Harrell, F. E., Lee, K. L., Tyor, M. P. and Pryor, D. B. 'Using ordinal logistic regression to estimate the likelihood of colorectal neoplasia', *Journal of Clinical Epidemiology*, **44**, 1263–1270 (1991).
10. Cox, C. 'Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach', *Statistics in Medicine*, **14**, 1191–1203 (1995).
11. Greenland, S. 'Alternative models for ordinal logistic regression', *Statistics in Medicine*, **13**, 1665–1677 (1994).
12. Hastie, T. J., Botha, J. L. and Schnitzler, C. M. 'Regression with an ordered categorical response', *Statistics in Medicine*, **8**, 785–794 (1989).
13. Koch, G. G., Amara, I. A. and Singer, J. M. 'A two-stage procedure for the analysis of ordinal categorical data', in Sen, P. K. (ed.), *BIOSTATISTICS: Statistics in Biomedical, Public Health and Environmental Sciences*, Elsevier Science Publishers B.V., North-Holland, 1985.
14. McCullagh, P. 'Regression models for ordinal data', *Journal of the Royal Statistical Society, Series B*, **42**, 109–142 (1980).
15. Peterson, B. and Harrell, F. E. 'Partial proportional odds models for ordinal response variables', *Applied Statistics*, **39**, 205–217 (1990).
16. Whitehead, J. 'Simple size calculations for ordered categorical data', *Statistics in Medicine*, **12**, 2257–2271 (1993).
17. Cole, T. J., Morley, C. J., Thornton, A. J., Fowler, M. A. and Hewson, P. H. 'A scoring system to quantify illness in babies under 6 months of age', *Journal of Royal Statistical Society, Series A*, **154**, 287–304 (1991).
18. Yee, T. W. and Wild, C. J. 'Vector generalized additive models', *Journal of the Royal Statistical Society, Series B*, **58**, 481–493 (1996).
19. Harrell, F. E., Lee, K. L., Califf, R. M., Pryor, D. B. and Rosati, R. A. 'Regression modelling strategies for improved prognostic prediction', *Statistics in Medicine*, **3**, 143–152 (1984).
20. Harrell, F. E., Lee, K. L. and Mark, D. B. 'Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors', *Statistics in Medicine*, **15**, 361–387 (1996).
21. MathSoft. *S-Plus User's Manual, Version 2.3*, MathSoft, Inc., Seattle WA, 1995.
22. Harrell, F. E. 'Design: S-Plus functions for biostatistical/epidemiologic modeling, testing, estimation, validation, graphics, prediction, and typesetting by storing enhanced model design attributes in the fit. UNIX and Microsoft Windows versions available from http://www.med.virginia.edu/medicine/clinical/hes/biostat.htm' 1997.

23. WHO/ARI Study Group on the Clinical Signs and Etiological agents of Pneumonia, Sepsis and Meningitis in Young Infants, 'Study methods', in preparation (1997).
24. Zhou, X. 'Effect of verification bias on positive and negative predictive values', *Statistics in Medicine*, **13**, 1737–1745 (1994).
25. WHO/ARI Study Group on the Clinical Signs and Etiological Agents of Pneumonia, Sepsis and Meningitis in Young Infants, 'Ordinal outcome scale', in preparation (1997).
26. Follmann, D. 'Multivariate tests for multiple endpoints in clinical trials', *Statistics in Medicine*, **14**, 1163–1175 (1995).
27. Harrell, F. E., Lee, K. L., Matchar, D. B. and Reichert, T. A. 'Regression models for prognostic prediction: Advantages, problems, and suggested solutions', *Cancer Treatment Reports*, **69**, 1071–1077 (1985).
28. Marshall, G., Grover, F. L., Henderson, W. G. and Hammermeister, K. E. 'Assessment of predictive models for binary outcomes: An empirical approach using operative death from cardiac surgery', *Statistics in Medicine*, **13**, 1501–1511 (1994).
29. D'Agostino, R. B., Belanger, A. J., Markson, E. W., Kelly-Hayes, M. and Wolf, P. A. 'Development of health risk appraisal functions in the presence of multiple indicators: The Framingham Study nursing home institutionalization model', *Statistics in Medicine*, **14**, 1757–1770 (1995).
30. Cureton, E. E. and D'Agostino, R. B. *Factor Analysis, An Applied Approach*, Erlbaum Publishers, New Jersey, 1983.
31. Sarle, W. S. 'The VARCLUS procedure', in *SAS/STAT User's Guide*, vol. 2, 4th edn, SAS Institute, Inc., Cary NC, 1990, Chapter 43, pp. 1641–1659.
32. Hoeffding, W. 'A non-parametric test of independence', *Annals of Mathematical Statistics*, **19**, 546–557 (1948).
33. Venables, W. N. and Ripley, B. D. *Modern Applied Statistics with S-Plus*, Springer-Verlag, New York, 1994.
34. Jolliffe, I. T. *Principal Component Analysis*, Springer-Verlag, New York, 1986.
35. Jackson, J. E. *A User's Guide to Principal Components*, Wiley, New York, 1991.
36. Kuhfeld, W. F., 'The PRINQUAL procedure', in *SAS/STAT User's Guide*, vol. 2, 4th edn, SAS Institute, Inc., Cary NC, 1990, Chapter 34, pp. 1265–1323.
37. Atkinson, A. C. 'A note on the generalized information criterion for choice of a model', *Biometrika*, **67**, 413–418 (1980).
38. Stone, C. J. and Koo, C. Y. 'Additive splines in statistics', *Proceedings of the Statistical Computing Section ASA*, 45–48 (1995).
39. Devlin, T. F. and Weeks, B. J. 'Spline functions for logistic regression modeling', in *Proceedings of the Eleventh Annual SAS Users Group International Conference*, SAS Institute, Inc., Cary NC, 1986, pp. 646–651.
40. Harrell, F. E., Lee, K. L. and Pollock, B. G. 'Regression models in clinical studies: Determining relationships between predictors and response', *Journal of the National Cancer Institute*, **80**, 1198–1202 (1988).
41. Herndon, J. E. and Harrell, F. E. 'The restricted cubic spline hazard model', *Communications in Statistics – Theory and Methods*, **19**, 639–663 (1990).
42. Lamport, L. *LaTeX: A Document Preparation System*, 2nd edn, Addison-Wesley, Reading, MA, 1994.
43. SAS Institute, Inc. *SAS/STAT User's Guide*, vol. 2, 4th edn, SAS Institute, Inc., Cary NC, 1990.
44. Schoenfeld, D. 'Partial residuals for the proportional hazards regression model', *Biometrika*, **69**, 239–241 (1982).
45. Grambsch, P. and Therneau, T. 'Proportional hazards tests and diagnostics based on weighted residuals', *Biometrika*, **81**, 515–526 (1994). Amendment and corrections in **82**, 668 (1995).
46. Cox, D. R. 'Regression models and life-tables (with discussion)', *Journal of the Royal Statistical Society Series B*, **34**, 187–220 (1972).
47. Landwehr, J. M., Pregibon, D. and Shoemaker, A. C. 'Graphical methods for assessing logistic regression models (with discussion)', *Journal of the American Statistical Association*, **79**, 61–83 (1984).
48. Collett, D. *Modelling Binary Data*, Chapman and Hall, London 1991.
49. Cleveland, W. S. 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American Statistical Association*, **74**, 829–836 (1979).

50. Altman, D. G. and Andersen, P. K. 'Bootstrap investigation of the stability of a Cox regression model', *Statistics in Medicine*, **8**, 771–783 (1989).
51. le Cessie, S. and van Houwelingen, J. C. 'Ridge estimators in logistic regression', *Applied Statistics*, **41**, 191–201 (1992).
52. Verweij, P. and van Houiwelingen, H. C. 'Penalized likelihood in Cox regression', *Statistics in Medicine*, **13**, 2427–2436 (1994).
53. Gray, R. J. 'Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis', *Journal of the American Statistical Association*, **87**, 942–951 (1992).
54. Hurvich, C. M. and Tsai, C. 'Regression and time series model selection in small samples', *Biometrika*, **76**, 297–307 (1989).
55. Schwarz, G. 'Estimating the dimension of a model', *Annals of Statistics*, **6**, 461–464 (1978).
56. Spiegelhalter, D. J. 'Probabilistic prediction in patient management', *Statistics in Medicine*, **5**, 421–433 (1986).
57. Tibshirani, R. 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society Series B*, **58**, 267–288 (1996).
58. Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J., *Classification and Regression Trees*, Wadsworth and Brooks/Cole Pacific Grove, CA, 1984.
59. Efron, B. and Tibshirani, R. *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.
60. Nagelkerke, N. J. D. 'A note on a general definition of the coefficient of determination', *Biometrika*, **78**, 691–692 (1991).
61. van Houwelingen, J. C. and le Cessie, S. 'Predictive value of statistical models', *Statistics in Medicine*, **8**, 1303–1325 (1990).
62. Phillips, A. N., Thompson, S. G. and Pocock, S. J. 'Prognostic scores for detecting a high risk group: Estimating the sensitivity when applied to new data', *Statistics in Medicine*, **9**, 1189–1198 (1990).
63. Brier, G. W. 'Verification of forecasts expressed in terms of probability', *Monthly Weather Review*, **75**, 1–3 (1950).
64. Faraway, J. J. 'The cost of data analysis', *Journal of Computational and Graphical Statistics*, **1**, 213–229 (1992).

TUTORIAL IN BIOSTATISTICS

Using observational data to estimate prognosis: an example using a coronary artery disease registry

Elizabeth R. DeLong[1,*,†], Charlotte L. Nelson[2], John B. Wong[3], David B. Pryor[4], Eric D. Peterson[2], Kerry L. Lee[5], Daniel B. Mark[2], Robert M. Califf[6] and Stephen G. Pauker[3]

[1]*Outcomes Research & Assessment Group, Duke Clinical Research Institute, Duke University, Department of Medicine, Biometry Division, Community and Family Medicine, 2400 Pratt Street, Durham, NC 27705, U.S.A.*
[2]*Outcomes Research & Assessment Group, Duke Clinical Research Institute, Duke University, Department of Medicine, Durham, NC 27710-7510, U.S.A.*
[3]*New England Medical Center, Department of Medicine, Box 302, Boston, MA 02111, U.S.A.*
[4]*Allina Health System, P.O. Box 9310, Minneapolis, MN 55440-9310, U.S.A.*
[5]*Duke Clinical Research Institute, Biometry Division, Community and Family Medicine, Box 3363, Durham, NC 27710-7510, U.S.A.*
[6]*Duke Clinical Research Institute, Duke University Division of Cardiology and Department of Medicine, Box 31123, Durham, NC 27710-7510, U.S.A.*

SUMMARY

With the proliferation of clinical data registries and the rising expense of clinical trials, observational data sources are increasingly providing evidence for clinical decision making. These data are viewed as complementary to randomized clinical trials (RCT). While not as rigorous a methodological design, observational studies yield important information about effectiveness of treatment, as compared with the efficacy results of RCTs. In addition, these studies often have the advantage of providing longer-term follow-up, beyond that of clinical trials. Hence, they are useful for assessing and comparing patients' long-term prognosis under different treatment strategies. For patients with coronary artery disease, many observational comparisons have focused on medical therapy versus interventional procedures. In addition to the well-studied problem of treatment selection bias (which is not the focus of the present study), three significant methodological problems must be addressed in the analysis of these data: (i) designation of the therapeutic arms in the presence of early deaths, withdrawals, and treatment cross-overs; (ii) identification of an equitable starting point for attributing survival time; (iii) site to site variability in short-term mortality. This paper discusses these issues and suggests strategies to deal with them. A proposed methodology is developed, applied and evaluated on a large observational database that has long-term follow-up on nearly 10 000 patients. Copyright © 2001 John Wiley & Sons, Ltd.

*Correspondence to: Elizabeth DeLong, Outcomes Research & Assessment Group, Duke Clinical Research Institute, Duke University, Department of Medicine, Biometry Division, Community and Family Medicine, 2400 Pratt Street, Durham, NC 27705, U.S.A.
†E-mail: delon001@mc.duke.edu

Contract/grant sponsor: Agency for Health Care Policy and Research; contract/grant numbers: HS08805, HS06503

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
 ISBN: 0-470-02365-1

1. INTRODUCTION

One of the most important aspects of clinical decision making is selecting treatment strategies for individual patients. A patient's general health, demographic status and disease severity will influence both the choice of therapy and the prognosis. For patients with coronary artery disease (CAD), the patient's risk profile can determine whether a more invasive, and costly, strategy is indicated. Because CAD accounts for a major portion of cardiovascular disease (the U.S.A.'s number one cause of death, with incurred costs exceeding $150 billion annually [1]), these decisions are relevant to society as well as to individual patients.

This study addresses some of the methodological issues in using observational data to create prognostic models for CAD patients under different therapeutic options. Currently, the three primary options after diagnosis by cardiac catheterization are medical therapy (MED), percutaneous transluminal coronary angioplasty (PTCA), and coronary artery bypass graft surgery (CABG), the latter two being revascularization procedures.

Ideally, data to assess the influence of treatment strategy on prognosis would come from randomized trials. However, these trials are expensive and often have limited sample size and follow-up. Further, they cannot always be generalized to the broader spectrum of patients and practice settings. Thus, observational data must sometimes supply information for medical decision-making. In this case, the choice of therapy, the prognosis, and the relative survival benefit depend to a great extent on the patient's risk profile. In particular, because treatment groups are not necessarily comparable prior to treatment, quantitative statistical models that attempt to account for treatment selection bias while estimating survival under alternative treatment strategies are needed. In addition to the selection bias that is inherent in treatment comparisons with such data, a number of other issues are relevant, and these additional issues are the focus of this manuscript.

1.1. Issues in assessing prognosis for CAD patients using observational data sources

A fundamental difficulty when using observational data to estimate prognosis for CAD patients involves defining treatment-specific survival. Physicians and patients who initially select a less invasive treatment option understand that later cross-over to a more invasive alternative is possible. For example, a patient may begin treatment with medical therapy, but later undergo CABG. The initial post-catheterization treatment 'strategy' incorporates this potential change in treatment. The prognosis that includes survival from treatment initiation through any subsequent cross-overs will be designated as arising from a 'treatment strategy' perspective. This perspective is in distinction to the 'single treatment' perspective, which evaluates prognosis while receiving only the initial treatment and censors at any cross-over, such as expected survival while on MED (described further in Section 1.2).

In randomized trials, treatment assignment is unbiased and survival time is initiated at randomization. Observational studies of CAD patients, however, often lack an explicitly recorded treatment assignment and thus have no uniformly logical treatment initiation time. Although the treatment decision occurs soon after the catheterization, it is generally not recorded in the observational data set. Furthermore, personal reasons or scheduling difficulties can delay the actual performance of a procedure for several days; consequently, some patients could be lost to follow-up if they go elsewhere for procedures or some may die while awaiting revascularization. For such patients, if no revascularization procedure has occurred following

catheterization, a default assignment to the medical therapy arm would attribute both early deaths and early losses to follow-up to MED. This policy assumes medical survival time begins at catheterization, whereas procedural survival begins at the procedure date. Using this convention and assuming the patient has survived the catheterization procedure, medical survival is unconditional, but procedural survival is implicitly conditioned on surviving to receive the procedure. This problem is similar to one described in transplantation literature. When patients who do not survive long enough to undergo transplantation are, in analyses, assigned to the non-transplant group, this creates a selection 'waiting time' bias in favour of the transplantation group. The methodological problem of when to begin 'survival time' has led to a serious and sustained debate questioning if heart transplantation survival benefits may have been caused by selection bias [2–5]. With the pace of care increasing, the delay from catheterization to PTCA is often far shorter than the delay from catheterization to CABG. Hence the potential 'waiting time bias' is particularly favourable toward CABG. This dilemma underscores the need for establishing an equitable 'treatment initiation' time.

A further issue in assessing prognosis is that peri-procedural care (usually defined as the first 30 days following a procedure) and long-term care are generally handled differently and by different types of care providers. The early survival after catheterization and treatment assignment is a variable component of overall survival that depends to a great extent on institutional constraints and individual care providers. This 'problem' also exists when physicians wish to make decisions based on RCT data. It is especially true for CABG and PTCA, which depend on the skill of the operator. Hence, local effects on early mortality need to be considered in long-term prognosis. Also, the factors that are important determinants of this early risk may differ from those affecting long-term survival.

An additional statistical concern is the differential risk associated with alternative treatments in the early period. CABG is known to incur a much higher early risk than either PTCA or MED, but this risk declines sharply in the first few days after the surgery. Because the early hazards for the three treatments are not proportional, a simple proportional hazards survival model cannot be used directly to obtain estimates of relative treatment hazards. Treatment effects in the later interval are more likely to conform to proportional hazards.

A survival model that allows the variable 30-day mortality component to be estimated independently (possibly locally) and then coupled with the long-term component may be used to compare prognosis under different short-term scenarios. In addition, estimates of the conditional survival (dependent on surviving this initial 30-day period) can be used by patients and providers following a successful procedure, or by those who want data that is not influenced by the early mortality rate.

1.2. Previous approaches

Some of the earliest observational prognostic assessments for CAD patients came from the CASS multi-site data registry [6, 7]. Acknowledging the difficulties of determining treatment assignment and exposure time, these studies performed analyses using several methods. One comparative analysis between MED and CABG assigned patients to medical therapy unless CABG was performed within an established time window. Patients who did not undergo CABG and either died or were lost to follow-up before the average time to CABG were excluded from analysis to avoid biasing the estimates against medical therapy survival. This

exclusion addressed the potential for waiting time bias, although the time to CABG is a skewed distribution with a heavy tail and a mean that is far greater than the median.

Another CASS method used what we designate as the 'single treatment' perspective for medical therapy, a method that was also used for survival comparisons in the Duke Cardiovascular Database [8–10]. In these comparisons, medical survival represented the interval of time a patient spent on medical therapy prior to death, revascularization or loss to follow-up. The medical survival of a patient who crossed over from medicine to a procedure was censored at the time of the procedure, and that patient's remaining survival was analysed as procedural survival. Whereas survival time for medical patients began at catheterization, the procedural survival times began at the time of procedure. Patients who died in the first few days after catheterization without undergoing a procedure, including those few deaths due to catheterization, were assumed to be medical failures.

Blackstone *et al.* [11] used parametric modelling of a cumulative hazard function to estimate survival to different time-related events following cardiac valve surgery. Their approach allows a time-varying decomposition of the hazard into as many as three phases, with the incorporation of potentially different covariates into each phase. Using this method, the problems of non-proportional early hazards and a separate covariate set for different time periods can be addressed directly. Treatment comparisons are not an issue in these analyses. Hence, because valve surgery has a logical treatment initiation and assignment, whereas medical therapy does not, this method does not address the issues of treatment assignment and initiation.

2. METHODS

2.1. Overview

We developed a strategy to define treatment assignment and treatment initiation and to accommodate a site- or region-specific 30-day mortality when using observational data for prognostic comparisons among CAD patients. This first step in our procedure was to implement a 'treatment strategy' approach by empirically designating a treatment assignment window. The next step was to designate a logical 'treatment initiation' point. Using the traditional 30-day period to define short-term mortality, we then separated the short- and long-term components of survival. Figure 1 demonstrates the proposed scheme for treatment assignment, treatment initiation and designation of the modelling population. Modelling efforts involved constructing a conditional survival model that would allow the short-term survival component for each of the three treatment strategies to be imported from an external source. The long-term conditional model was thus estimated on the population of all patients who initiated treatment and then survived the short-term 30-day period after treatment initiation. This framework allows for site- or region-specific short-term mortality models to be combined with the long-term conditional survival model to estimate overall prognosis for individuals or classes of patients.

2.2. Application population

We developed and applied our methods using the Duke Cardiovascular Database. This prospectively collected database and its follow-up procedures have been described elsewhere [12]. The patient population for the current study was 9251 consecutive patients referred to Duke University Medical Center for cardiac catheterization between 1 March 1984 and 1 August 1990.

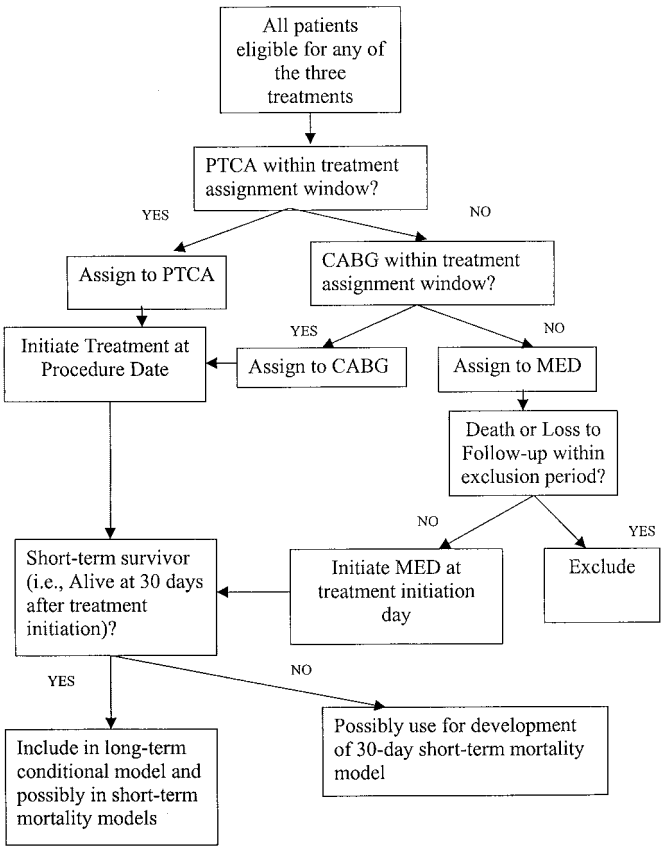


Figure 1. Decision rule for treatment assignment, treatment initiation, and inclusion in modelling population.

Patients were included if they had not previously undergone a revascularization procedure and were considered eligible for all three of the treatment options. This latter criterion excluded patients with left main disease, because they generally undergo CABG. Follow-up extended to July 1994.

2.3. Treatment perspective

In a 'treatment strategy' perspective, the initial treatment decision is considered a strategy that incorporates the opportunity for subsequent cross-over to an alternative treatment. For example, a patient who lives one year on medical therapy and an additional five years following CABG is considered to have lived six years after the decision to begin with medical therapy. Hence, medical survival represents the survival that could be attributed to a strategy of initial medical therapy ('intention to treat'), with subsequent revascularization possible depending on clinical circumstances.

We used Kaplan–Meier [13] survival curves to display differences in treatment assignment perspectives. It will be informative to recall that the Kaplan–Meier curve is constructed as a product over all event times; at any particular event time, the contribution is a fraction that contains the population at risk in the denominator and the number of non-events in the numerator. For purposes of comparison, we calculated Kaplan–Meier [13] survival curves for patients assigned to MED according to three different perspectives designated as 'single treatment', 'treatment strategy', and 'medicine only'. With the 'single treatment' perspective, every patient who survives the catheterization is assigned to the MED group, if only temporarily. Survival time is initiated at catheterization and extends until the patient has an event (death) while still on medical therapy; it is censored when the patient is lost to follow-up or undergoes a procedure (PTCA or CABG), in which case the treatment assignment transfers to the respective procedure. On the other extreme, the 'medicine only' treatment perspective defines the MED group to include only patients who never underwent revascularization during follow-up; their survival time is censored at loss to follow-up. This latter group of patients is a subset of the 'single treatment' MED group and includes a mixture of patients who are not considered healthy enough to withstand a procedure, along with those who are considered not to need a procedure. The events (deaths) in these two groups are identical because they include all deaths among patients who have never been revascularized. Hence, any differences between the two Kaplan–Meier curves are due to differences in the populations at risk at each event time, which are reflected in the denominators of the components of the curves.

The 'treatment strategy' perspective provides an intermediate definition of the MED group, but requires that a treatment assignment window be designated. For our comparisons, we implemented three different treatment assignment windows, as described below.

2.4. Treatment assignment window

As noted previously, the intended treatment assignment is rarely captured in a computerized observational database. Hence, an initial goal of these analyses was to establish a treatment assignment time window. The treatment assignment window represents a period of time following cardiac catheterization, during which patients who were initially intended to receive a procedure would likely receive it. Within this window, patients who underwent procedures are assigned to the respective revascularization strategy (patients who underwent both procedures are assigned to that which occurred first). All other patients, including those who received PTCA or CABG after the treatment assignment window, are assigned to the MED strategy. Thus, the MED population at risk for a 'treatment strategy' perspective includes all of the patients identified by the 'medicine only' perspective. Within the treatment assignment window, the risk set (denominator) for the Kaplan–Meier curve at any event time is a subset of the 'single treatment' MED at-risk population. The numerator is identical, because all events are among patients not yet revascularized in each case. Beyond the treatment assignment window, the risk sets and events for the 'treatment strategy' perspective include patients who underwent late procedures.

To determine an appropriate treatment assignment window, we first evaluated the cumulative distributions of days until CABG for all patients in the database who eventually underwent a CABG procedure and days until PTCA for all patients who eventually underwent a PTCA procedure. Based on these distributions and current clinical practice, we selected 30, 45 and 60 days following catheterization as candidates for a treatment assignment window. We then

compared the Kaplan–Meier survival curves for the medically assigned patients under each of these potential treatment assignment windows (see Section 3) and also compared them with the 'medicine only' group and the 'single treatment' group, defined above.

2.5. Treatment initiation and waiting time bias

With no explicit treatment assignment recorded in an observational data record, the assignment of early deaths to a treatment strategy is problematic. To address the problem of potential waiting time bias, which can adversely affect comparative MED survival estimates, we created a 'treatment initiation' designation for medical therapy. Patients who had not undergone a revascularization procedure and either died or were lost to follow-up prior to the 'treatment initiation' point (a number of days after catheterization) were excluded. This was an attempt to equalize the starting point for survival across treatments, so that medical survival can be considered conditional on 'treatment initiation' in the same way that procedural survival is conditional on undergoing the procedure.

To assess the impact of assigning early deaths to the medical therapy arm and to find a reasonable 'treatment initiation' time point, we inspected the distribution of times to early death or loss to follow-up (LTF) for patients assigned to medical therapy according to the designated treatment assignment window. We then compared the MED Kaplan–Meier survival curves for three candidate 'treatment initiation' points of 0, 7 and 14 days.

CABG and PTCA treatment strategies were initiated at procedure date. The starting time points (0.0 on the survival time axis) for all three strategies were then conditional on treatment initiation. At treatment initiation, the survival time clock began and survival was defined to be 100 per cent at this point for all three strategies.

2.6. Data configuration for long-term conditional survival model

The long-term survival modelling relied on Cox regression analysis to account for patient characteristics while assessing the effects of treatment strategy on an underlying, unspecified, hazard function. Because there are no published short-term medical mortality models, our particular application allows for external CABG and PTCA short-term models, but uses the Duke Database to estimate the MED short-term mortality. Other institutions may have their own short-term MED models that reflect their local experience and procedures, including admission criteria, use of specialty services, nursing protocols and transitional care. They may wish to supplement their short-term data with a long-term prognostic model for which they do not have local data.

We thus separated the short-term modelling from the long-term modelling component. The first module (called the short-term module) comprised all patients who initiated therapy and their 30-day survival was either imported or modelled separately. The second module (called the long-term conditional module) consisted of all patients who survived at least 30 days after treatment initiation, regardless of treatment assignment. Thus, the short-term module represents the initial 30-day survival, conditioned only on treatment initiation. The long-term conditional module provides a long-term component that is conditional on 30-day survival for each of the three treatment strategies. Figure 2 displays a hypothetical modelling scenario using this configuration.

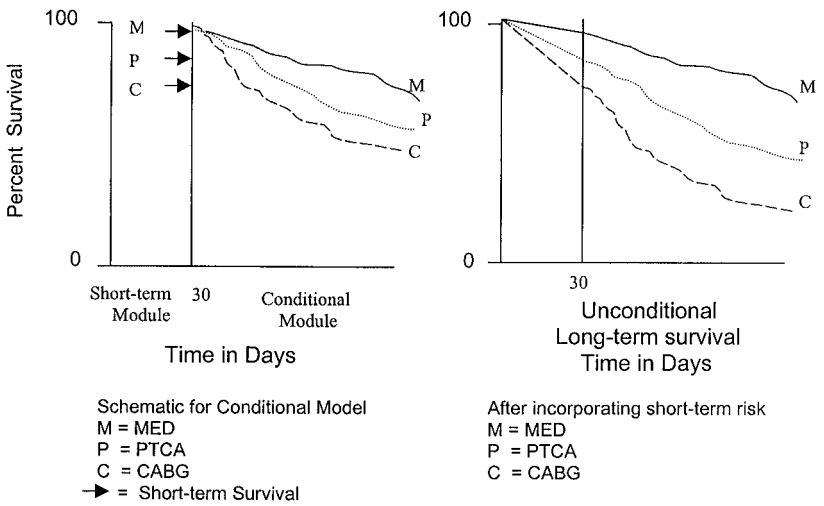


Figure 2. Schematic of hypothetical conditional model and subsequent overall survival model.

We implemented the long-term module by modelling the underlying conditional hazard as

$$\lambda(t \mid X_i) = \lambda_0(t) \exp(\beta X_i + \beta_P I_P + \beta_C I_C), \quad t \geqslant 30$$

where $\lambda(t \mid X_i)$ is the underlying hazard at time t for a patient with covariate vector X_i. I_P and I_C are indicator variables for PTCA and CABG, respectively, and β_P and β_C are the corresponding parameter estimates. The underlying hazard, $\lambda_0(t)$, starts at day 30 and applies to all patients surviving at least 30 days after treatment initiation.

This modelling structure incorporates some basic assumptions. First, a patient's specific hazard is assumed to relate to the underlying hazard by the proportionality factor, $\exp(\beta X_i)$. The model is also parameterized so that the PTCA and CABG strategies are each represented by indicator covariates and are thus assumed to conform to the proportional hazards assumption in the conditional stratum. This assumption can be tested, as described below. An additional assumption is that the MED long-term conditional module absorbs the short-term procedural mortality incurred by the small percentage of patients who are assigned to medical therapy but later cross over from MED to a procedure. In addition, although our application involves a single site, other databases may accumulate long-term data from several sites; this strategy can be employed in those databases with the addition of site-specific parameters.

2.7. Model construction

Standard model building procedures were used to select covariates for the long-term survival model, to test for non-proportional hazards, and to determine significant variable interactions. Our analytic approach was to test model variables first for significance, and then for violations of the proportional hazards assumption. Variable transformations were implemented during the testing phase to help improve the model fit and to resolve violations of model assumptions.

The proportional hazards assumption was tested by assessing the log hazard ratio over time using scaled Schoenfeld partial residuals [14, 15]. We also used time-dependent covariates to test for time by predictor interactions. For the few covariates that violated the proportional hazards assumption, we stratified the model by estimating different underlying hazards rather than hazard ratios (taking into consideration that hazards are not proportional across strata, and hence an estimated hazard ratio is not appropriate). The resulting model can be written as

$$\lambda^k(t \mid X_i) = \lambda_0^k(t) \exp(\beta X_i + \beta_P I_P + \beta_C I_C), \quad t \geqslant 30$$

where k indexes the covariate strata.

2.8. Adequacy of the conditional model

We used several measures of observed versus expected survival to assess the fit of the conditional model in various subgroups. Because the underlying hazard is allowed to vary across strata, model estimates will be likely to fit the individual strata well. We therefore evaluated the adequacy of the model in subgroups that had not been used to stratify the model and that included patients from more than one stratum. The most relevant of these were the three subgroups determined by the number of diseased vessels (1, 2 or 3). For each of these subgroups, we generated the observed (Kaplan–Meier) five-year survival curve along with the average individual conditional predicted survival at each failure time point.

The long-term conditional model was used to estimate individual patient survival curves up to five years. As described above, the estimates from this module represent long-term conditional survival for all three-treatment strategies. This conditional survival is estimated as

$$\hat{S}_X^k(t) = \hat{S}_0^k(t)^{\exp(\hat{\beta}X_i + \hat{\beta}_P I_P + \hat{\beta}_C I_C)}, \quad t \geqslant 30$$

for a patient in stratum k with covariate vector X_i. $\hat{S}_0^k(t)$ is the underlying estimated survival curve derived by iteratively solving a maximum likelihood equation involving the estimated parameters from the proportional hazards model and a set of hazard contributions at each failure time [16]. It is produced by most software programs that perform Cox survival modelling.

For each subgroup, we averaged the individual survival predictions over all patients in the subgroup (for example, one vessel disease). The resultant average curve represents the estimated truncated survival at five years for this subgroup. Several comparative measures of model fit were then generated. First, the average observed survival at five years, calculated from the Kaplan–Meier curve, was compared with the average estimated five-year survival. We also calculated three discrepancy measures between the overall curves. The first is the maximum absolute difference between the curves over the entire five years, expressed as per cent survival. This measure reflects the maximum error at any time point. The second measure is the difference between the areas under the two curves, which can be interpreted as an estimate of difference in five-year life expectancy between observed and predicted. The third measure is the total absolute area between the two curves, as a measure of 'absolute deviation'. This last measure can be envisioned as the sum of areas of all of the regions that are defined by the two curves. It is identical to the difference between the areas under the two curves when one curve always lies above the other, but the two measures can differ substantially if the curves cross.

2.9. Computing overall survival and testing for differences between treatments

Unconditional survival from 30 days forward can be calculated for all three treatment strategies by multiplying the long-term conditional survival by (1–30-day mortality), using the respective short-term mortality. To complete the survival curve from day 0 (treatment initiation) to day 30, the survival at time 0 (1.0) can be linearly connected to the 30-day survival estimate.

As an example of the short-term CABG risk, we used the in-hospital mortality model developed by Hannan *et al.* [17] in the New York State CABG Surgery Reporting System data. For PTCA mortality, we applied the model developed in the Cooperative Cardiovascular Project [18]. MED short-term mortality was modelled within the Duke Database. Using these estimates, a five-year truncated life expectancy for an individual patient can be calculated as the area under the curve for each of the three treatment strategies. This set of triplets (one for each patient) can used to investigate whether there are certain groups of patients for whom treatment differences are maximized on this response scale.

3. APPLICATION

3.1. Determining the treatment assignment window

Of the 9251 study patients, 6506 eventually underwent a PTCA or CABG during follow-up. Cumulative distributions of time from catheterization to CABG, PTCA and either procedure for these patients are displayed in Figure 3. For 2855 of these patients, the initial procedure was PTCA, with 75 per cent undergoing PTCA within 4 days and 91 per cent within

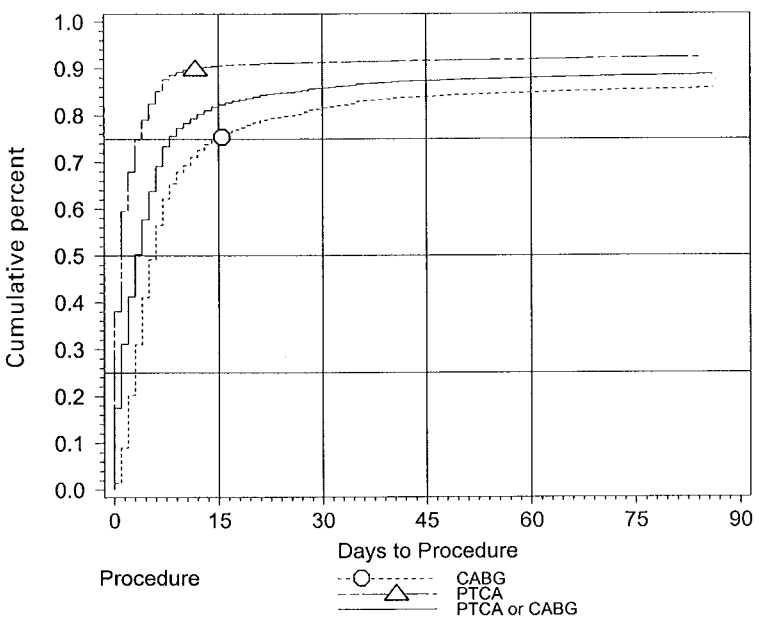


Figure 3. Cumulative distributions time to procedure for PTCA, CABG and either.

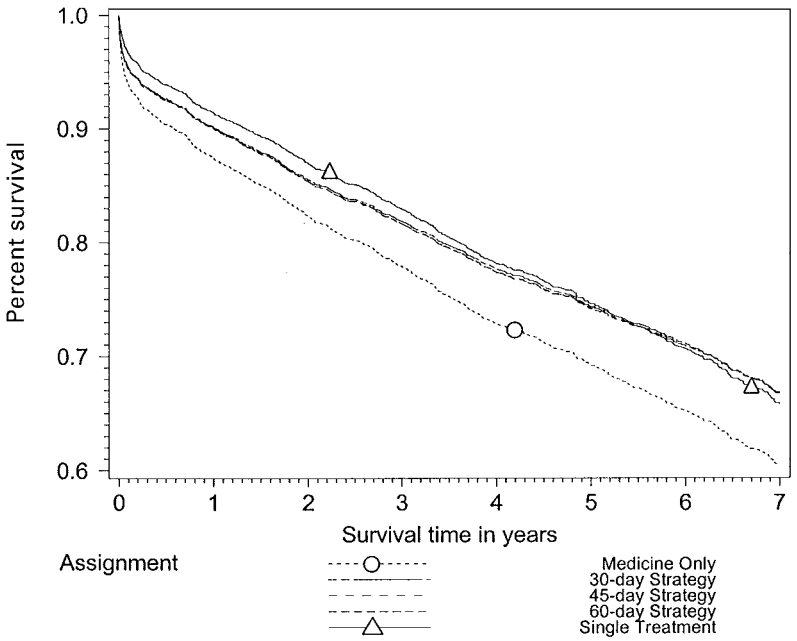


Figure 4. Kaplan–Meier survival comparison among methods for assigning patients to MED.

30 days of catheterization. An additional 3651 patients underwent CABG as a first procedure, 75 per cent within 14 days and 81 per cent within 30 days of catheterization. The overall 75th percentile for time to procedure for CABG and PTCA patients combined was 8 days. By 60 days after catheterization, all three cumulative distributions appear to level off, indicating that after this period, new procedures accumulate slowly. In fact, the distributions are extremely skewed: the 90th percentile for time to PTCA is 12 days and the 95th percentile is 466 days; the 90th percentile for time to CABG is 487 days and the 95th percentile is 1311 days.

Figure 4 demonstrates the seven-year Kaplan–Meier survival comparison among the 'treatment strategy', the 'medicine only', and the 'single treatment' approaches for assignments to the medical therapy arm. At this point in the analysis, no early deaths have been excluded from any of the groups. The 'single treatment' MED group includes all patients until they undergo a procedure, are lost to follow-up, or die without having a procedure. The MED group defined by the 30-day treatment assignment window is initially a subset of the 'single treatment' group and contains all patients except those who undergo procedures within 30 days of catheterization. Likewise, the 45- and 60-day groups form decreasing subsets of the 'single treatment' group. The 'medicine only' group is the limiting extension of the treatment assignment window and only includes patients who never underwent a procedure. In the first 30 days after catheterization, the Kaplan–Meier curves can only differ due to the sizes of the populations at risk (denominators), because no procedural deaths are counted as events in any of these groups. After 30 days, the 30-day treatment assignment group also contains

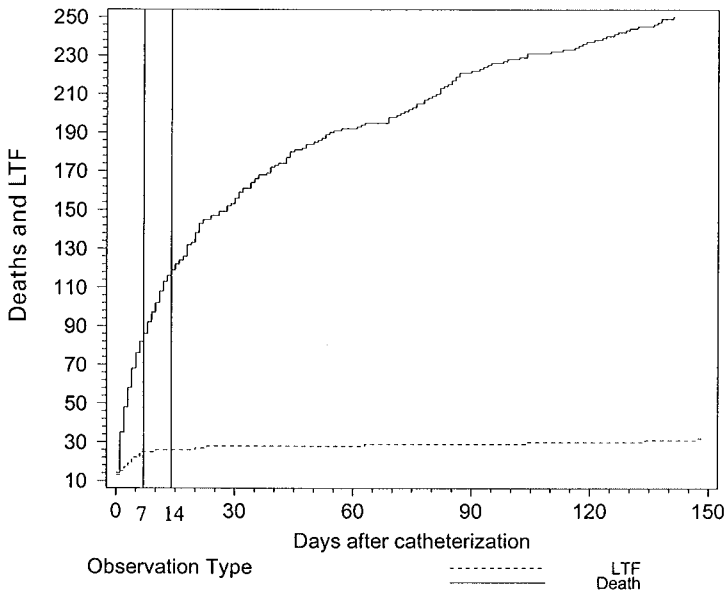


Figure 5. Cumulative deaths and losses to follow-up (LTF) without undergoing PTCA or CABG.

events among those few patients who undergo procedures after 30 days and subsequently die during follow-up. Similar reasoning holds for the 45- and 60-day groups. The estimated 'single treatment' survival has a substantial early advantage over the other curves because of the inflated denominator in the Kaplan–Meier calculations.

As expected, the 'medicine only' curve is substantially below the others, in part because patients who are never referred for procedure are generally a higher-risk group. This phenomenon is also seen when the 'single treatment' curve crosses the 'treatment strategy' curves at about five years, indicating that the healthier patients are being censored from the 'single treatment' curve in the later years. Patients who are considered hardy enough to withstand a procedure are selectively removed from the medical 'single treatment' group, leaving the sicker patients [19]. These healthier patients are retained in the medical group when the 'treatment strategy' perspective is taken.

Among the three candidates for treatment assignment windows (30, 45 and 60 days), there is virtually no difference in the survival curves. Hence, because 30 days incorporates over 90 per cent of all PTCA procedures, 80 per cent of all CABG procedures, and 85 per cent of all procedures in general, it was selected for a logical treatment assignment window.

3.2. Determining the treatment initiation point

Using the 30-day treatment assignment window for defining the MED arm, Figure 5 displays the cumulative distribution of early deaths and losses to follow-up (LTF) for medically assigned patients. There are a large number of each during the first 7 days, but after that the LTF are infrequent and the number of deaths continues to accumulate, but less dramatically.

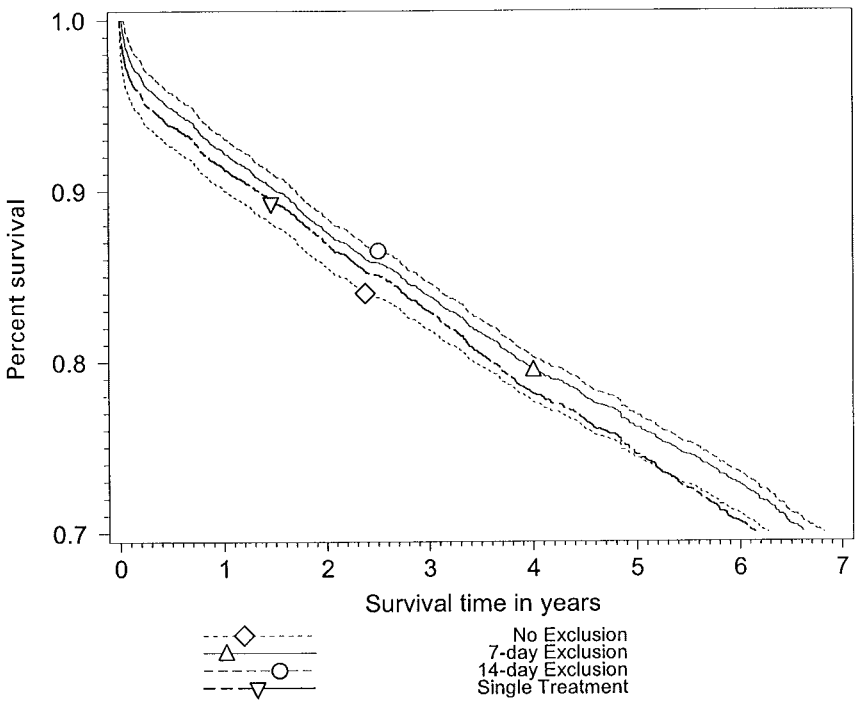


Figure 6. Kaplan–Meier survival comparisons varying exclusion period for determining MED treatment initiation.

Figure 6 compares the Kaplan–Meier survival curves computed by using no exclusions, or 7 or 14 days as exclusions for a 'treatment initiation' period. For reference, the 'single treatment' curve is also plotted.

We reason that the survival of patients actually intended to receive medical therapy is compromised when the early failures and LTF who are not intended for medical therapy are included. Thus, the actual medical survival curve would not be expected to drop as precipitously in the early period as the 'single treatment' curve, which includes the deaths of all patients who are not revascularized. Both the 7- and 14-day exclusion periods remedy this situation. Synthesizing the information from Figures 5 and 6, we selected the 7-day exclusion period as a reasonable compromise.

Thus, based on clinical reasoning and empirical data, we established 30 days as the 'treatment assignment' window and the eighth day following catheterization as the 'treatment initiation' point. In Figure 1, which displays the decision algorithm for assigning patients in the survival model, the treatment assignment window is now determined as 30 days, not to be confused with the short-term mortality period, which is also 30 days. The exclusion period is now 7 days in Figure 1. Only patients who remained alive at least a week after catheterization without being revascularized can be considered to have initiated medical therapy. For example, a patient who died 15 days after catheterization without undergoing a procedure was assigned to the medical treatment strategy and had an exposure time of 8 days of survival

after treatment initiation. A patient who underwent CABG at 35 days post catheterization and died 9 days after that was also considered to be on the medical treatment strategy (because the patient crossed over to surgery more than 30 days after the catheterization) and was credited with 37 $(35 + 9 - 7)$ days of medical survival.

Of the 2855 patients undergoing PTCA as the first procedure after catheterization, 2606 had their procedures within 30 days and were assigned to the PTCA group. Similarly, 2978 patients were assigned to the CABG group. After excluding 86 early deaths and 25 LTF, an additional 3556 who were assigned to medical therapy initiated medical therapy. The 673 medical patients who underwent CABG and the 249 who underwent PTCA more than 30 days after catheterization comprised a cross-over rate from MED of 19 per cent to CABG and 7 per cent to PTCA.

For the conditional modelling, we excluded 99 deaths within 30 days of CABG and 100 PTCA deaths within 30 days. This PTCA mortality rate of 3.8 per cent was relatively high and is due to the large number of patients with myocardial infarction receiving PTCA at this tertiary care institution. (Other institutions may have a less critically ill population with better estimates for low risk patients.) In addition, 82 of the 3556 patients who initiated medical therapy (2.3 per cent) died in the first 30 days following treatment initiation.

3.3. Long-term conditional survival modelling

We determined that both MI status and extent of disease were significantly important prognostic factors for conditional survival and that neither conformed to the proportional hazards assumption. After stratifying on three levels of MI status (1 = MI within 24 hours, 2 = recent MI within 6 weeks or currently unstable, 3 = no recent MI) and two levels of disease severity (1 = one or two vessel disease with no proximal left anterior descending disease; 2 = two vessel disease with proximal LAD or three vessel disease), no further violations of proportional hazards were encountered (recall that patients with left main disease were excluded because they were not considered candidates for all three treatment options). In this conditional model, treatment effects were modelled directly, without violating the proportional hazards assumptions.

Besides treatment effects, the final model included the extent of coronary artery disease, as coded by the CAD index (range 23 to 74), incorporating the number and location of diseased vessels [9, 12] ejection fraction, age, gender, severity of congestive heart failure (CHF), mitral regurgitation (MR), and a Charlson [20] index that was modified to exclude myocardial infarction and CHF. The relation between ejection fraction and survival was different for medical patients than for patients who underwent CABG or PTCA. We also found that the effect of increasing CAD index was different for the three treatments so we included interaction terms to capture these differential relationships.

3.4. Adequacy of the conditional model

Figures 7(*a*)–(*c*) display the results of the modelling process. The computer code for generating a typical curve is given in Example 1 of the appendix. For PTCA and CABG, mean predicted conditional survival are plotted against the observed Kaplan–Meier survival for one, two and three vessel patients. For MED, because we created both short- and long-term models within the Duke data set, we plot the entire unconditional curves. Table I tabulates the accompanying survival estimates and discrepancy measures. The conditional survival curves

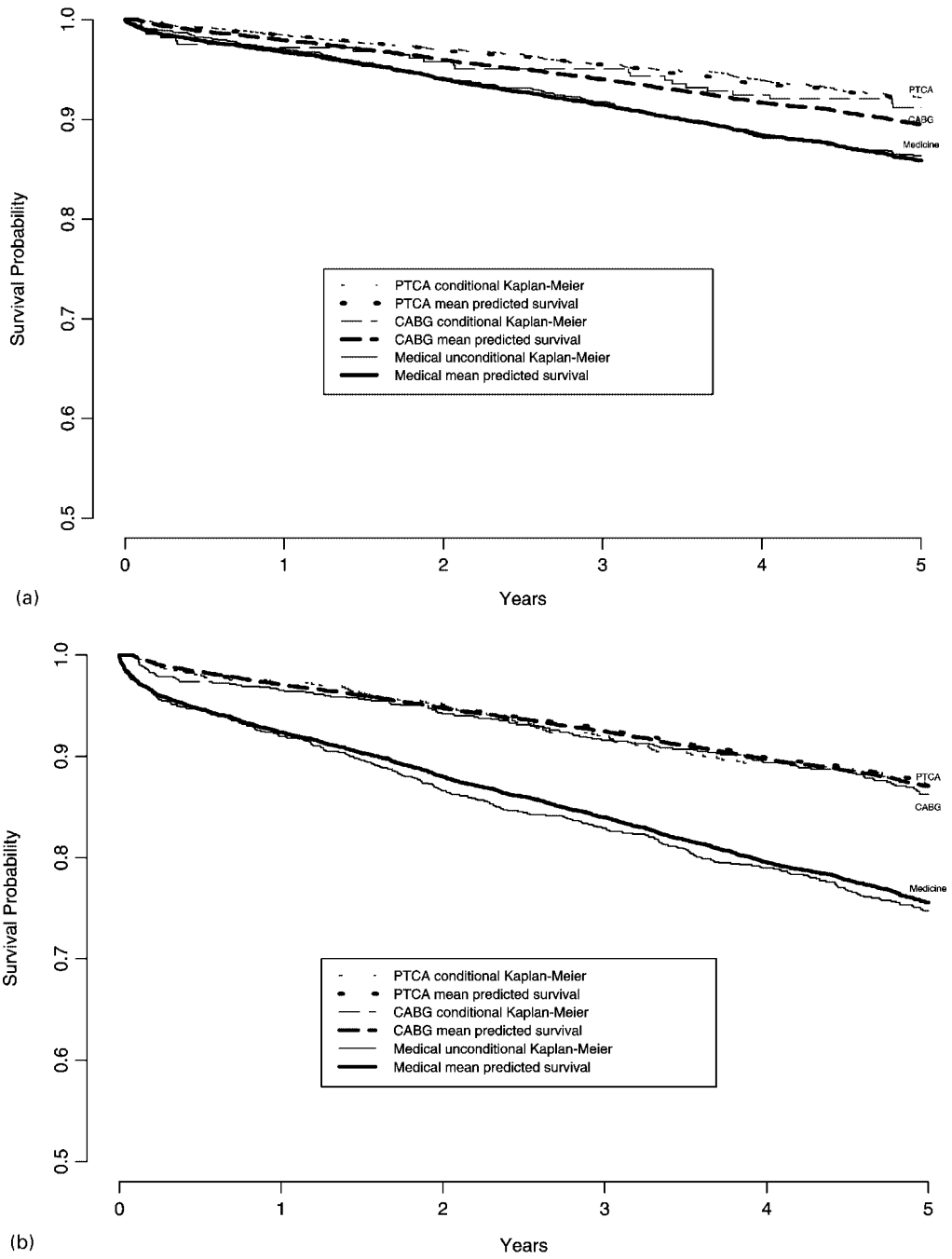


Figure 7. Observed and predicted survival: (*a*) single-vessel disease; (*b*) two-vessel disease; (*c*) three-vessel disease.

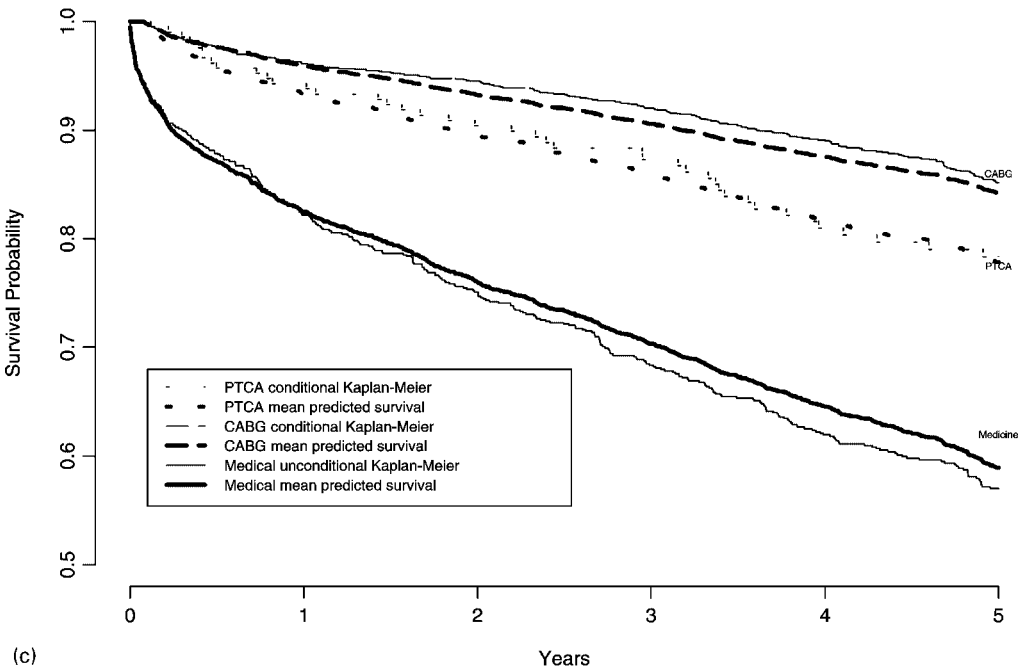


Figure 7. Continued.

Table I. Five-year survival estimates and measures of discrepancy for the final conditional model.

Subgroup	*N*	Survival estimates at five years		Discrepancy measures		
		Kaplan–Meier	Predicted	Maximum difference* (%)	Difference between areas†	Absolute deviation‡
One vessel						
Medicine	1669	0.863	0.859	0.5	0.007	0.010
PTCA	1569	0.922	0.920	0.6	0.007	0.008
CABG	284	0.912	0.895	2.0	0.009	0.037
Two vessel						
Medicine	1088	0.747	0.755	1.7	0.041	0.041
PTCA	728	0.872	0.875	1.4	0.015	0.021
CABG	979	0.863	0.870	1.3	0.024	0.024
Three vessel						
Medicine	799	0.570	0.586	2.9	0.062	0.071
PTCA	209	0.783	0.778	2.0	0.029	0.040
CABG	1616	0.851	0.842	1.6	0.046	0.047

*Difference in per cent survival over five years.
†Five year difference between areas under Kaplan–Meier and predicted curves.
‡Five year total area between the curves.

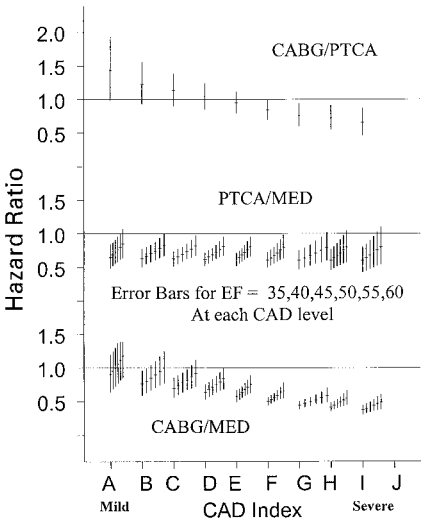


Figure 8. Ninety five per cent confidence intervals for hazard ratios. Ejection fraction from 35 to 60 within each CAD index level for PTCA/MED and for CABG/MED.

begin at 100 per cent at day 30. The scale of these plots is from 50 per cent to 100 per cent survival probability to magnify areas of discrepancy. For single vessel disease, the maximum difference between observed and expected curves is 2.0 per cent at 4.75 years for CABG; for MED and PTCA, the maximum differences are 0.5 per cent and 0.6 per cent, respectively. For two-vessel disease, the maximum discrepancy of 1.7 per cent occurs at 2.5 years for the medically treated group, and is 1.4 per cent at 3 months for CABG and 1.3 per cent at 2.5 years for PTCA. The prediction for three-vessel MED patients demonstrates the greatest discrepancy from observed survival with a maximum difference of 2.9 per cent at 4 years. Maximum differences for PTCA and CABG were 2.0 per cent and 1.6 per cent, respectively. The difference in five-year truncated life expectancy is less than 0.05 years (less than 20 days), with an absolute deviation less than 0.05, for all subgroups except the medically treated three-vessel disease group. As expected, the discrepancy measures tend to be largest in those subgroups for which five-year survival is worst, such as medically treated patients with three-vessel disease.

The conditional treatment hazard ratios vary with level of CAD index and ejection fraction. Hence, we calculated them at each level of CAD index for 5-unit increments of ejection fraction from 35 to 60. These are displayed in Figure 8 along with 95 per cent confidence intervals, with CAD index increasing in severity from the first level (level A) to the highest level (level J). (The computer code for producing this figure is given in Example 2 of the appendix.) Because of the finding that the CABG/PTCA hazard ratios do not depend on ejection fraction, there is a single confidence interval for this comparison for each value of CAD index. For the other two comparisons, as ejection fraction increases from 35 to 60, the superiority over MED becomes less pronounced. For this long-term component of survival, without discounting for procedural mortality, PTCA confers significantly lower risk than medical therapy at essentially all levels of CAD index and ejection fraction, except for

Table II. Conditional and unconditional truncated life expectancy over five years for each treatment strategy.

Treatment	Conditional truncated survival Mean years ± standard deviation	Unconditional truncated survival Mean years ± standard deviation
Medicine	4.35 ± 0.63	4.36 ± 0.71
CABG	4.56 ± 0.38	4.53 ± 0.45
PTCA	4.52 ± 0.45	4.53 ± 0.51

marginal superiority at the highest ejection fraction values. CABG has increasing superiority over medical therapy with more severe coronary artery disease. For patients at the highest levels of CAD index, CABG is also significantly superior to PTCA.

Although CABG appears to be the treatment of choice for most patients when conditional survival is evaluated, the overall five-year truncated life expectancy depends on the expected procedural mortality. We first calculated the conditional truncated life expectancy over five years for all three treatments by taking the areas under the individual patient estimated curves. Then, using the Hannan CABG procedural mortality model [17], the CCP PTCA [18] procedural mortality model, and the internally developed MED short-term model, we also calculated the corresponding unconditional figures. Table II displays the means and standard deviations of these measures. Whereas CABG yields somewhat better conditional survival in this group of patients, the PTCA and CABG unconditional survivals are essentially equivalent after accounting for the differential short-term mortality. Because of the steep early mortality for CABG, the unconditional survival, obtained by multiplying the conditional curve by the 30-day CABG survival and then adding the 30-day component, turns out to be less than the conditional survival. This result implies that, once the patient has survived 30 days from procedure, life expectancy is greater than it was prior to the surgery.

Figure 9 demonstrates the effect of incorporating a variable procedural mortality into the long-term treatment decision. Note that the scale of the survival probability axis has been truncated at 0.7 as a lower bound, to highlight differences in the curves. This figure displays estimated five-year survival for a typical patient treated with PTCA with 1.5 per cent acute mortality risk versus CABG with 1.5 per cent operative mortality risk and CABG with a 4 per cent operative mortality risk. Here, five-year survival is clearly superior for CABG when acute mortality is 1.5 per cent, but when CABG acute mortality is 4 per cent, patient preferences with respect to risk aversion and long-term versus short-term benefit may play a significant role.

4. DISCUSSION

We have used an example from coronary artery disease to demonstrate issues in the use of observational data, for both a treatment comparison analysis and for supplying information to the medical decision making process. We used a ‘treatment strategy’ approach to develop a statistical model that assesses long-term survival for patients with coronary artery disease following an initial treatment decision among MED, PTCA and CABG. Using a data framework that creates a treatment assignment and attributes survival conditional on treatment initiation,

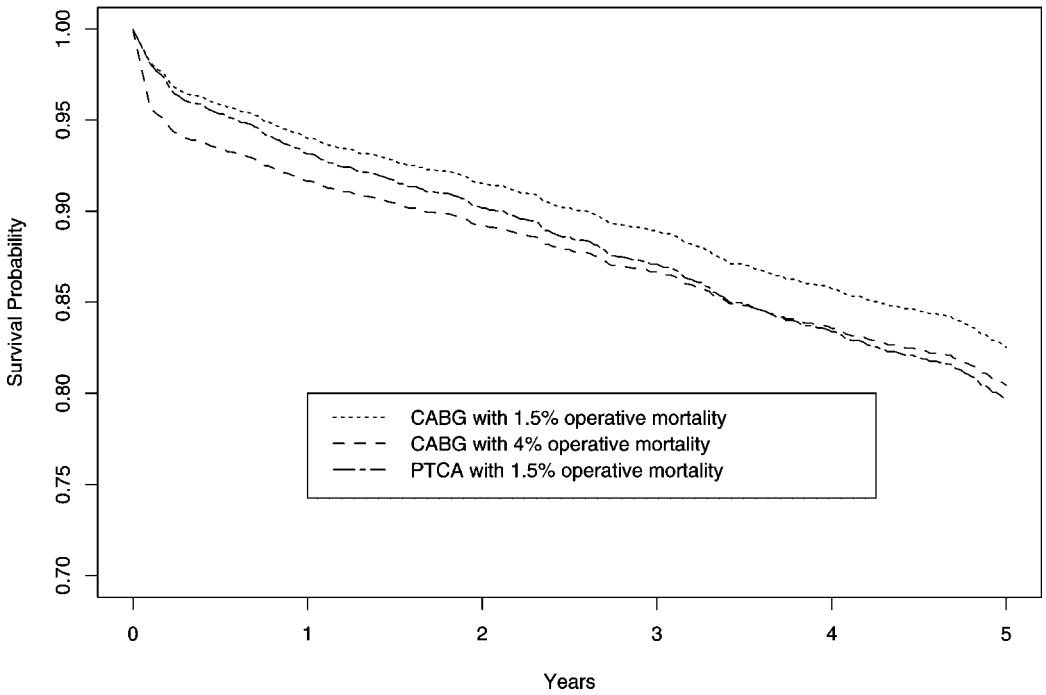


Figure 9. Predicted long-term survival as a function of acute risk.

the model allows for independently developed short-term mortality components. Our approach simulates an 'intention to treat' analysis because survival is credited to the treatment initially selected. It also establishes a context for the analyses. Whereas survival after PTCA or CABG is implicitly conditional on having lived to receive the procedure, survival on MED has no corresponding initiation event.

Peduzzi and colleagues [21], in analysing the VA randomized trial of CABG versus MED, demonstrated that a randomized trial can be subject to severe bias if it does not adhere to the 'as randomized' treatment assignment. Their study compared four different approaches to the 'as randomized' analysis of randomized data and concluded that when observational analyses: (i) credit long waiting times for surgery to the surgery group; or (ii) exclude patients who cross over from medicine to surgery; or (iii) count a surgical death after a long medical follow-up as early surgical mortality, they are likely to incur bias. They concluded that the gold standard 'as randomized' analysis may not, in fact, compare actual treatments, but rather it compares treatment 'strategies'. This was precisely the goal of the present analyses, which incorporate methods to avoid the above sources of bias and produce a survival model for treatment strategy, rather than treatment received.

Our 'treatment strategy' perspective for treatment comparisons attributes survival after changes in treatment to the initial assignment. Other perspectives would require different analytic approaches and possibly different interpretations. For example, another perspective might assess survival on the ultimate treatment, rather than the initial treatment. Such analyses would need to carefully assess the implicit waiting time bias prior to an intervention. The

problem has been addressed in the transplantation literature, in which several parametric and non-parametric approaches for comparing survival among transplanted and non-transplanted patients have been tried [3] including the use of time-dependent covariates [4]. Some analyses have proposed a modification of comparative life tables to incorporate the transient states transplantation candidates traverse [5]. No single approach has been advocated as a mechanism for solving the waiting time bias problem.

A multiple failure time perspective could also be employed [22], whereby PTCA and CABG would be considered as different types of failures and any individual could experience multiple events. We chose not to take this perspective because the occurrence of PTCA or CABG is influenced to a great extent by the practising clinician. In addition, our goal was to compare treatment strategies with respect to survival, rather than to evaluate the time until treatment.

Our analysis offers a mechanism for dealing with some specific issues with regard to estimating prognosis from observational data. We are not proposing a particular method for the problem of treatment selection bias, which is an inherent issue in most observational analysis. We used a risk adjustment model in our application, although other methods are available. For example, Rosenbaum and Rubin [23] demonstrate that propensity scores can be effective in estimating treatment differences, assuming treatment assignment and treatment initiation have been decided.

The treatment comparison results from our model are consistent with previously published prognostic studies for CAD patients. After attempting to account for treatment selection, we found that for less severe presentation of disease, PTCA is superior to medicine. For the most severe presentation of disease, CABG is superior, recognizing that CABG may not be an option for patients with significant comorbid burden. For intermediate disease states, the treatment strategy cannot be determined on the basis of long-term survival, but must also consider patient preferences and local procedural success rates, which can have a substantial impact on the overall prognosis.

One advantage of our conditional approach is that it allowed us to incorporate and estimate the effects of treatment strategy directly in the model, without stratifying on this variable. After determining that the hazard associated with a treatment strategy varies with CAD index and ejection fraction, we were able to estimate these relative hazards. With a model that begins survival time at the time of procedure, proportional hazards are clearly violated because of the procedural mortality risk.

In addition, this approach optimizes the clinical utility of both long- and short-term data collection efforts. Few sites have the resources to accumulate long-term follow-up on large numbers of patients. However, short-term procedural outcome data are increasingly available. Risk algorithms that accurately account for early mortality following these procedures have been extensively studied and are in the public domain. The ability to account for or vary the early mortality in assessing long-term prognosis is an appealing aspect for medical decision making.

APPENDIX

Below are examples of computer code used to generate two of the figures presented in this manuscript. The first is an S-plus program that plots observed Kaplan–Meier curves overlaid

on the average estimated patient-specific survival function. The second is an SAS program for calculating confidence intervals for odds ratios and hazard ratios.

Example 1: note comments and documentation in S-plus begin with a #

```
##########################################################################
# Plot observed and predicted survival curve                             #
# Input:                                                                 #
#   km.object - A Kaplan-Meier object from a model fit using cph,        #
#               stratified on treatment (TREAT) and number of            #
#               diseased vessels (NO).                                   #
#   my.data   - An Splus dataframe containing all data to be used        #
#               for prediction.                                          #
#   my.model  - The Cox Proportional Hazards survival model; an          #
#               object fit using cph.                                    #
##########################################################################

library(Design,T) ######################################################
                  # Attach Frank Harrell's Design library.               #
                  # Harrell FE (2000): Design: S functions for           #
                  # biostatistical/epidemiologic modeling, testing,      #
                  # stimation, validation, graphics, and prediction.     #
                  # Programs available from lib.stat.cmu.edu or          #
                  # hesweb1.med.virginia.edu/biostat/s/Design.html.      #
                  ########################################################

store()
postscript(horizontal=T)

########## Obtain vector of time points for predicted survival     ####
##########          from 0.0 to 5.0 years                          ####
timepoints<-seq(0, 5,by =.02)
last.time<-length(timepoints)
pmv(last.time)


##############      PTCA, NUMBER OF DISEASED VESSELS = 3        #######

########    Subset dataset on PTCA treated, no=3 subset  ##############
dfptca<-my.data[my.data$treat=="PTCA" & my.data$no==3,]

######## Use survival model to predict survival for each patient ######
########         in this subset at each specified timepoint      ######
pred.ptca <- survest.cph(my.model,dfptca,times=timepoints,conf.int=F)

######## Calculate mean survival estimate at each timepoint      ######
```

```
mean.ptca<-apply(pred.ptca$surv,2,mean)
last.ptca<-mean.ptca[last.time]
pmv(last.ptca)


###############      CABG, NUMBER OF DISEASED VESSELS = 3        #######

######## Subset dataset on CABG treated, no=3 subset ##################
dfcabg<-my.data[my.data$treat=="CABG" & my.data$no==3,]

######## Use survival model to predict survival for each patient ######
########         in this subset at each specified timepoint      ######
pred.cabg <- survest.cph(my.model,dfcabg,times=timepoints,conf.int=F)

######## Calculate mean survival estimate at each timepoint      ######
mean.cabg<-apply(pred.cabg$surv,2,mean)
last.cabg<-mean.cabg[last.time]
pmv(last.cabg)



###############      MED, NUMBER OF DISEASED VESSELS = 3        ########

######## Subset dataset on Med treated, no=3 subset ###################
dfmed<-my.data[my.data$treat=="Med" & my.data$no==3,]

######## Use survival model to predict survival for each patient ######
########         in this subset at each specified timepoint      ######
pred.med <- survest.cph(my.model,dfmed,times=timepoints,conf.int=F)

######## Calculate mean survival estimate at each timepoint      ######
mean.med<-apply(pred.med$surv,2,mean)
last.med<-mean.med[last.time]
pmv(last.med)


####### Observed and Conditional predicted suvival NO=3 Plots #########
survplot(km.object,treat="PTCA",no=3,ylim=c(.5,1),xlim=c(0,5),
         lty=8,label.curves=F,adj.subtitle=F,n.risk=F,pr=T)

lines(pred.ptca$time,mean.ptca,lty=8,lwd=3)

survplot(km.object,treat="CABG",no=3,ylim=c(.5,1),xlim=c(0,5),
         lty=4,label.curves=F,adj.subtitle=F,n.risk=F,add=T,pr=T)

lines(pred.cabg$time,mean.cabg,lty=4,lwd=3)
```

```
survplot(km.object,treat="Med",no=3,ylim=c(.5,1),xlim=c(0,5),
         lty=1,label.curves=F,adj.subtitle=F,n.risk=F,add=T,pr=T)

lines(pred.med$time,mean.med,lty=1,lwd=3)


legend(1.25,.75,c("PTCA conditional Kaplan-Meier",
                  "PTCA mean predicted survival",
                  "CABG conditional Kaplan-Meier",
                  "CABG mean predicted survival",
                  "Medical conditional Kaplan-Meier",
                  "Medical mean predicted survival"),cex=.75,
       lty=c(8,8,4,4,1,1),lwd=c(1,3,1,3,1,3))
title("Observed and Predicted Conditional Survival
Three vessel disease",cex=.8)
mtitle()

text(5,last.ptca+.02,"PTCA",cex=.5)
text(5,last.cabg+.02,"CABG",cex=.5)
text(5,last.med-.02,"Medicine",cex=.5)
##########################################################################
```

Example 2: note comments and documentation begin with an asterisk and end with a semicolon

```
*This macro uses output estimates from SAS PROC LOGISTIC
or SAS PROC PHREG to create confidence intervals for treatment
odds ratios (LOGISTIC) or hazard ratios (PHREG) as a function of
other covariates when the model contains treatment by covariate
interactions. The macro is customized to a specific PHREG model
that has two treatment indicators (PTCA CABG) compared against the
treatment MED, one covariate that interacts with these variables,
CADINDEX, and one covariate (EJECFRAFC) that interacts such that the
effect is the same for PTCA and CABG, but differs from that of MED.
The interaction variables are CADPTCA (CADINDEX*PTCA), CADCABG (
CADINDEX*CABG), and MEDEJEC (MED*EJECFRAC). The macro outputs a
dataset for plotting the estimated confidence intervals as a function
of CADINDEX and EJECFRAC;

%macro OR_RR ;
*Prepare dataset to be used by IML by keeping only the covariates that
will be used;
data params; set params;
*Keep only the parameters that will be used to calculate the odds
ratios or hazard ratios;
  keep ptca cabg cadptca cadcabg medejec;
```

```
  if upcase(_name_) in ('SURVTIME', 'PTCA', 'CABG', 'CADPTCA',
'CADCABG', 'MEDEJEC');

  proc iml workspace=50;
  use params;
  read all into total [colname=cols];

*Separate the model parameter estimates from their covariance matrix;
  covmax=total[2:nrow(total),];
  meanvec=total[1,];

*Set range of values for CADINDEX;
   cadindex={23 32 37 42 48 56 63 67 74};
*Set range of values for medical ejection fraction: MEDEJEC;
   medejec={35 40 45 50 55 60};

*Do each combination in turn;
   do jj=1 to ncol(medejec);
      ef=medejec[,jj];
      do ii=1 to ncol(cadindex);
        cad1=cadindex[,ii];

*Calculate the estimates for the PTCA vs MED odds or risk ratio;
  *First create linear combination vector for PTCA vs MED;
        cadx_vec=cad1||{0};
        ef\_vec=-ef;  *The parameter is associated with MED, so the
negative is needed for PTCA;
        ptca_vec={1 0} ||cadx_vec||ef_vec;
  *Calculate the point estimate of the PTCA/MED odds or risk ratio;
        ptca_mn=ptca_vec* meanvec';
  *Find the covariance matrix;
        ptca\_var=ptca\_vec*covmax*ptca_vec';
        ptca_se=sqrt(ptca_var);
  *Create confidence intervals;
        ptca_low=ptca_mn -1.96 *ptca_se;
        ptca_hi =ptca_mn + 1.96* ptca_se;

*Repeat for CABG vs MED;
        cadx_vec={0}||cad1;
        cabg_vec={0 1} ||cadx_vec||ef_vec;
  *Calculate the point estimate of the CABG/MED odds or risk ratio;
        cabg_mn=cabg_vec* meanvec';
  *Find the covariance matrix;
        cabg_var=cabg_vec*covmax*cabg_vec';
        cabg_se=sqrt(cabg_var);
```

```
  *Create confidence intervals;
        cabg_low=cabg_mn - 1.96*cabg_se;
        cabg_hi =cabg_mn + 1.96*cabg_se;


  *Now do the CABG vs PTCA ratios - these do not vary with ejection
fraction;
        cadx_1=-cad1;
        cadx_vec=cadx_1||cad1;
        cp_vec={-1 1}||cadx_vec||{0};
  *Calculate the point estimate of the CABG vs PTCA odds or risk ratio;
        cp_mn=cp_vec*meanvec‘;
  *Find the covariance matrix;
        cp_var=cp_vec*covmax*cp_vec‘;
        cp_se=sqrt(cp_var);
  *Create confidence intervals;
        cp_low=cp_mn-1.96*cp_se;
        cp_hi=cp_mn+1.96*cp_se;


*Concatenate all estimates into a vector to be output to a dataset and
exponentiate;
      vector= ptca_mn||ptca_se||ptca_low||ptca_hi
            ||cabg_mn||cabg_se||cabg_low||cabg_hi
            ||  cp_mn||  cp_se||  cp_low||  cp_hi;
      vector=exp(vector);
  *Add values of ejection fraction and cadindex;
      vector=ef||cad1||vector;
      if ii + jj=2 then holding=vector;
      else holding=holding//vector;
         end;
         end;


*Create variable names such that ptca refers to PTCA vs MED, cabg
refers to
           CABG vs MED and cp refers to CABG vs PTCA;
 namespec={‘ejecfrac’ ‘cadindex’ ‘ptca_mn’ ‘ptca_se’ ‘ptca_lo’
‘ptca_hi’ ‘cabg_mn’ ‘cabg_se’ ‘cabg_lo’ ‘cabg_hi’ ‘cp_mn’ ‘cp_se’
‘cp_lo’ ‘cp_hi’ };
 run;


*Output the dataset;
  create dplots from holding [colname=namespec];
  append from  holding;
```

```
   quit;
   %mend;


*Run the stratified survival model with PROC PHREG;

proc phreg data=itt2   covout outest=params  ;
model survtime*dead(0)=chfindex age50 mitral charlson sex cadindex
                      ptca cabg  ejecfrac cadptca cadcabg medejec
                     /ties=efron; strata  migrp cadigp;
run;

*Call the macro to calculate confidence intervals as a function of
CADINDEX and EJECFRAC;

%or_rr;


proc sort data=dplots; by cadindex ejecfrac;

data dplots; set dplots; by cadindex ejecfrac;

*offset each successive level of EJECFRAC by .7, so that they can be
embedded within the CADINDEX range;
    retain offset;
    if first.cadindex then offset=0;
       else offset+.7;
   cad_ef=cadindex+offset;
*Keep only the lowest level of EJECFRAC for the CABG vs PTCA comparison
because the hazard ratio does not vary with EJECFRAC;
   if ejecfrac>35 then do; cp_mn=.; cp_hi=.; cp_lo=.; end;
*Change the vertical axis to separate the sets of confidence intervals.
NOTE:  the axis labels for the resulting plot will have to be changed
manually;
     else do; cp_mn=cp_mn+4; cp_hi=cp_hi+4; cp_lo=cp_lo+4; end;
     ptca_mn=ptca_mn+2; ptca_hi=ptca_hi+2; ptca_lo=ptca_lo+2;

    cp=cp_hi; cm=cabg_hi; pm=ptca_hi; output;
    cp=cp_mn; cm=cabg_mn; pm=ptca_mn; output;
    cp=cp_lo; cm=cabg_lo; pm=ptca_lo; output;


*Set options for the plot;
goptions hsize=6in vsize=8in;
goptions device=cgmmwwc gsfname=gsfile gsfmode=replace hpos=40 vpos=40;
  filename gsfile 'FILENAME';   ********REPLACE WITH FILENAME;
```

```
*Set symbol and axis specifications;
symbol1 i=hilo v=none color=black;
symbol2 i=hilo v=none color=black;
symbol3 i=hilo v=none color=black;
axis2  offset=(1 cm)  order=0 to 6 by .5 major = (height=.7)
        label=(height=1.0 r=90 a=-90 "Hazard Ratio")
        minor=none;
axis1 minor=(width=1) offset=(1 cm) order=23 32 37 42 48 56 63 67 74
80;

*Call PROC GPLOT, using vref to separate panels;
proc gplot data=dplots;
plot  cp*cad_ef=1 cm*cad_ef=2 pm*cad_ef=3 /overlay vref=1  3 5
vaxis=axis2 haxis=axis1;
      label cad_ef= 'CAD Index';
  run;
```

REFERENCES

1. American Heart Association. *Heart And Stroke Facts: A Statistical Supplement*. American Heart Association, 1997.
2. Gail MH. Does cardiac transplantation prolong life? *Annals of Internal Medicine* 1972; **76**:815–817.
3. Turnbull BW, Brown BW, Hu M. Survivorship analysis of heart transplant data. *Journal of the American Statistical Association* 1974; **69**:74–80.
4. Crowley J, Hu M. Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association* 1977; **72**:27–36.
5. Mantel N, Byar DP. Evaluation of response-time data involving transient states: an illustration using heart-transplant data. *Journal of the American Statistical Association* 1974; **69**:81–86.
6. Chaitman BR, Davis K, Fisher LD, Bourassa MG, Mock MB, Lesperance J, Rogers WJ, Fray D, Tyras DH, Judkins MP, Ringqvist I, Killip T, participating CASS Hospitals. A life table and Cox regression analysis of patients with combined proximal left anterior descending and proximal left circumflex coronary artery disease: non-left main equivalent lesions (CASS). *Circulation* 1983; **68**:1163–1170.
7. Alderman EL, Fisher LD, Litwin P, Kaiser GC, Myers WO, Maynard C, Levine F, Schloss M. Results of coronary artery surgery in patients with poor left ventricular function (CASS). *Circulation* 1983; **68**:785–795.
8. Califf RM, Harrell FE Jr, Lee KL, Rankin JS, Hlatky MA, Mark DB, Jones RH, Muhlbaier LH, Oldham HN, Pryor DB. The evolution of medical and surgical therapy for artery disease: a 15-year perspective. *Journal of the American Medical Association* 1989; **261**:2077–2086.
9. Mark DB, Nelson CL, Califf RM, Harrell FE Jr, Lee KL, Jones RH, Fortin DF, Stack RS, Glower DD, Smith LR, DeLong ER, Smith PK, Reves JG, Jollis JG, Tcheng JE, Muhlbaier LH, Lowe JE, Phillips HR, Pryor DB. Continuing evolution of therapy for coronary artery disease: initial results from the era of coronary angioplasty. *Circulation* 1994; **89**:2015–2025.
10. Jones RH, Kesler K, Phillips HR III, Mark DB, Smith PK, Nelson CL, Newman MF, Reves JG, Anderson RW, Califf RM. Long-term survival benefit of coronary artery bypass grafting and percutaneous transluminal angioplasty in patients with coronary artery disease. *Journal of Thoracic Cardiovascular Surgery* 1996; **111**:1013–1025.
11. Blackstone EH, Naftel D, Turner M Jr. The decomposition of time-varying hazard into phases, each incorporating a separate stream of concomitant information. *Journal of the American Statistical Association*, 1986; **81**:615–624.
12. Smith LR, Harrell FE Jr, Rankin JS, Califf RM, Pryor DB, Muhlbaier LH, Lee KL, Mark DB, Jones RH, Oldham HN, Glower DD, Reves JG, Sabiston DC Jr. Determinants of early versus late cardiac death in patients undergoing coronary artery bypass graft surgery. *Circulation* 1991; Suppl III **84**(5):245–253.
13. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; **53**:457–481.
14. Schoenfeld DA. Partial residuals for the proportional hazards regression model. *Biometrika* 1982; **69**:239–241.
15. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994; **81**:515–526.

16. Kalbfleisch J, Prentice R. *The Statistical Analysis Of Failure Time Data*. Wiley: New York, 1980.
17. Hannan EL, Kilburn H Jr, Racz M, Shields E, Chassin MR. Improving the outcomes of coronary artery bypass surgery in New York State. *Journal of the American Medical Association* 1994; **271**:761–766.
18. Peterson ED, DeLong ER, Muhlbaier LH, Rosen AB, Fortin DF, Mark DB, Kiefe CI, Kresowik TF, Jencks SF, Pryor DB. Predicting mortality following coronary angioplasty: results from the Cooperative Cardiovascular Project. *Circulation* 1995; **92**:I–476.
19. Pryor DB, Lee KL, Harris PJ, Harrell FE Jr, Rosati RA. The effect of crossovers on estimates of survival in medically treated patients with coronary artery disease. *Journal of Chronic Diseases* 1984; **37**:521–529.
20. Charlson ME, Ales KL, Simon R, MacKenzie R. Why predictive indexes perform less well in validation studies: Is it magic or methods? *Archives of Internal Medicine* 1987; **147**:2155–2161.
21. Peduzzi P, Wittes J, Detre K, VA Cooperative Studies Program, Holford, T. Analysis as-randomized and the problem of non-adherence: an example from the veterans affairs randomized trial of coronary artery bypass surgery. *Statistics in Medicine* 1993; **12**:1185–1195.
22. Wei LJ, Glidden DV. An overview of statistical methods for multiple failure time data in clinical trials. *Statistics in Medicine* 1997; **16**:833–839.
23. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.

Part III
CLINICAL TRIALS

3.1 Design

TUTORIAL IN BIOSTATISTICS
DESIGNING STUDIES FOR DOSE RESPONSE

WENG KEE WONG

UCLA Department of Biostatistics, Los Angeles, CA 90024-1772, U.S.A.

AND

PETER A. LACHENBRUCH

FDA/CBER/OELPS HFM-215, 1401 Rockville Pike, Rockville, MD 20852, U.S.A.

SUMMARY

'Dose response' refers to the regression of a response on a stimulus. We review a number of options for dose–response designs, and compare various designs which may be used in practice. We start with two group designs. Next, we introduce basic optimal approximate design theory for simple linear and quadratic regression illustrating different criteria of optimality and their effect on the allocation of the levels of the dose. Then we obtain the efficiencies of these optimal approximate designs and some simple designs which have intuitive appeal (symmetry, equal spacing of treatments, reduced numbers of observations at the highest and lowest doses).

1. INTRODUCTION

The regression of response on stimulus may be represented graphically as a curve [as in Figure 1]. When the stimulus is in the form of a 'dose' (e.g., of a drug, or possibly of an applied force or some other source), this may be called a 'dose response curve'. (Kotz and Johnson[1]). In its simplest form, a dose–response curve is a simple linear or polynomial regression. More complex 'dose–response' curves may involve, for example, a transcendental function. Others may involve transformations of the dose in the regressions. For example, dose–response models often use the logarithm of dose. This function of the dose is called the dose metameter. In some cases, the response is quantal (yes/no) and the dose–response technique is a probit analysis or logit analysis. Figure 1 shows dose–response curves for linear and quadratic models.

The determination of a threshold dose is also a dose–response problem (see Figure 1). Here the response is A below a dose x_0 and B above that point. That is, the model is $E(Y|x) = A$ if $x < x_0$ and $E(Y|x) = B$ if $x \geqslant x_0$ where $E(Y|x)$ is the expected value of the response Y given $X = x$. The model requires that we estimate the value of x_0, A and B. In this tutorial paper we will consider the dose–response cases in which response and the dose are continuous and the regression functions are either simple linear or quadratic models, that is

$$E(Y|x) = A + Bx$$

or

$$E(Y|x) = A + Bx + Cx^2.$$

We also assume throughout that x is coded so that $0 \leqslant x \leqslant 1$.

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
 ISBN: 0-470-02365-1

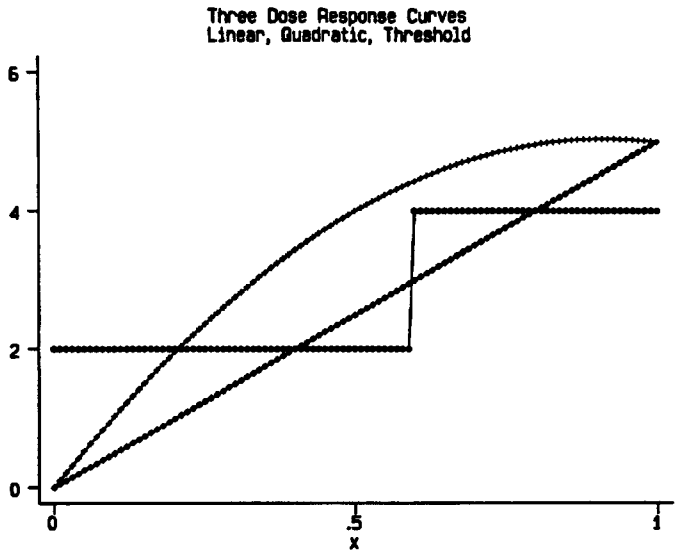


Figure 1. Three dose response curves: linear, quadratic and threshold

The steps in conducting a dose–response study (in idealized form) consist of the following:

(a) Assume a form or model for the curve (for example, linear, quadratic or threshold).
(b) Select the dose metameter (dose or log (dose)).
(c) Design the study so 'good' information is obtained; this includes obtaining estimates of the model coefficients with small standard errors and having the ability to test for model failures (such as testing for a quadratic model when a linear one has been assumed or testing for non-normal errors).
(d) Collect the data.
(e) Perform the analyses. For the simplest models, these include a linear regression, followed by model diagnostics such as testing for common variance (homoscedasticity), normality, and whether the model is linear or quadratic, and examining for outliers.
(f) Prepare a report describing the steps in the study, including the limitations of the study.

The objective of this paper is to discuss design issues in a dose–response experiment. Specifically, we consider the problem of allocating the dose, x, in $[0, 1]$ to estimate $E(Y|x)$. We motivate these issues by simulating data from three designs and two dose–response functions. Each data set has 20 observations with a standard deviation of 2. In all cases, the data are normally distributed. The goal is to estimate the dose–response function, which may be linear or quadratic. The first design allocates half of the data at $x = 0$ and half at $x = 1$. The second design has one-quarter of the observations at four points, $x = 0$, $x = 1/3$, $x = 2/3$ and $x = 1$. These are examples of *uniform* designs which have equal spacing of the x values, and equal number of observations at each x. The third design has half of the data at $x = 0$ and half at $x = 0{\cdot}75$. This design might be used if the investigator were concerned over possible toxic effects of the highest dose (at $x = 1$). The allocation of data is shown as histograms in Figure 2. The first response function is $E(Y|x) = 5x$. The second response function is $E(Y|x) = 11x - 6x^2$. This function was chosen to reach a maximum inside the interval $[0, 1]$. It also agrees with the linear function at 0 and 1. The data are given

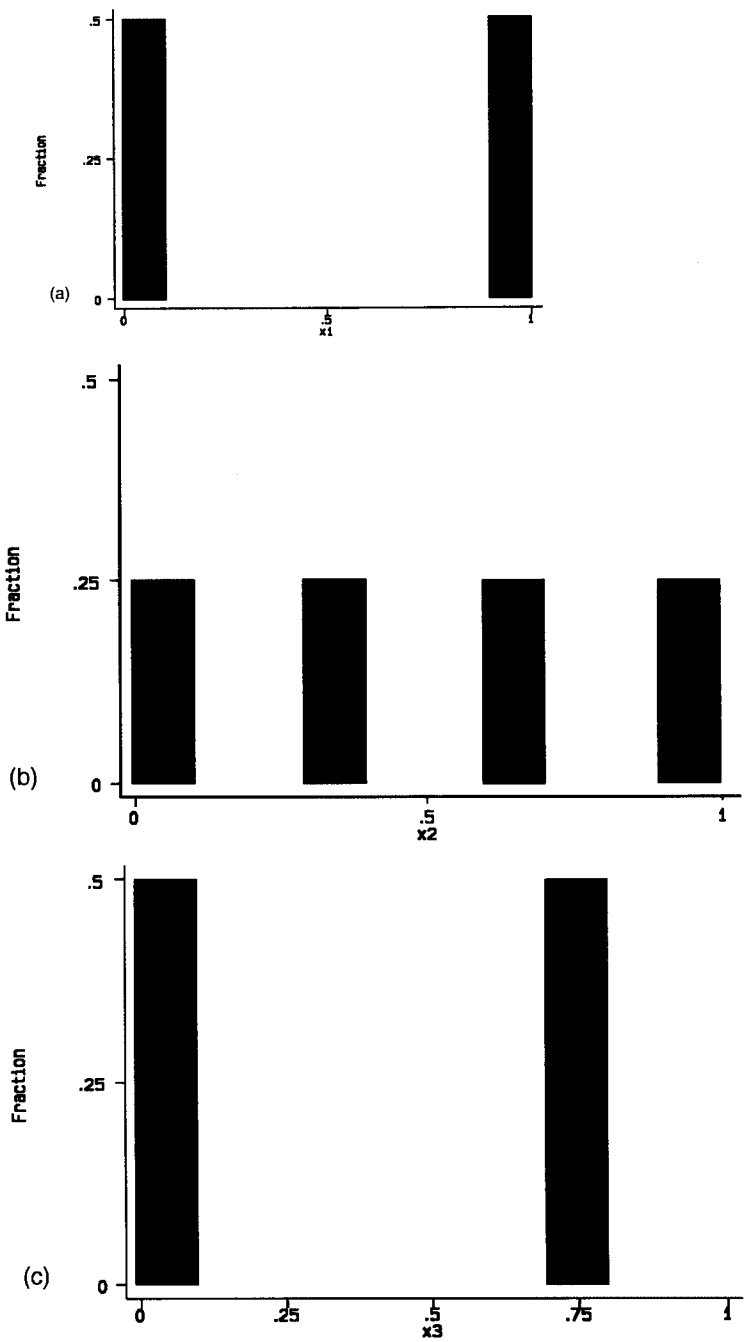


Figure 2. Histogram of X values for three designs

in Table I(a) and I(b) rounded by 5 digits (your analyses with these data should be similar, but not identical to ours). To distinguish the data of Table I(a) from that of Table Ib) we label the simulated dependent variable data in Table Ib) as W.

Table I. Data and analysis from three simulated dose–response studies

(*a*) Data generated by $Y_i = 5x_i + \varepsilon$, where ε is normal with $\mu = 0$ and $\sigma = 2$

	Design 1		Design 2		Design 3	
	x_1	y_1	x_2	y_2	x_3	y_3
1	0	− 3·69894	0	− 1·77029	0	− 3·95725
2	0	0·88581	0	1·04965	0	1·42738
3	0	4·24483	0	− 0·94450	0	0·45690
4	0	0·66185	0	0·99879	0	4·77256
5	0	− 1·37427	0	2·35421	0	− 1·48124
6	0	2·54575	0·33	4·84129	0	1·40287
7	0	− 0·84730	0·33	2·03856	0	1·28237
8	0	− 3·48385	0·33	0·47213	0	0·86857
9	0	− 0·97389	0·33	4·40601	0	− 0·98921
10	0	1·55282	0·33	0·29701	0	1·68386
11	1	8·21102	0·67	2·39918	0·75	1·52494
12	1	2·27225	0·67	5·94551	0·75	0·65396
13	1	7·51096	0·67	3·13616	0·75	8·59038
14	1	6·38050	0·67	7·11803	0·75	2·62081
15	1	6·58215	0·67	− 0·64908	0·75	5·79675
16	1	2·86575	1	1·60837	0·75	7·02660
17	1	7·94210	1	4·43520	0·75	4·30115
18	1	1·05613	1	3·62980	0·75	5·18883
19	1	8·61141	1	3·16162	0·75	4·42332
20	1	3·90077	1	3·44190	0·75	5·01722

(*b*) Data generated by $W_i = 11x - 6x^2 + \varepsilon$, where ε is normal with $\mu = 0$ and $\sigma = 2$

	Design 1		Design 2		Design 3	
	x_1	w_1	x_2	w_2	x_3	w_3
1	0	− 1·77029	0	− 3·69894	0	− 0·92930
2	0	1·04965	0	0·88581	0	− 0·36553
3	0	− 0·09445	0	4·24483	0	− 0·27144
4	0	0·99879	0	0·66185	0	− 1·70517
5	0	2·35421	0	− 1·37427	0	1·43117
6	0	3·19129	0·33	5·52235	0	− 2·25067
7	0	0·38856	0·33	2·12930	0	− 0·53015
8	0	− 1·17787	0·33	− 0·50725	0	− 1·74972
9	0	2·75601	0·33	2·00270	0	− 2·18116
10	0	− 1·35299	0·33	4·52942	0	3·14768
11	1	4·04918	0·67	7·88762	0·75	5·37605
12	1	7·59551	0·67	1·94885	0·75	2·13208
13	1	4·78616	0·67	7·18757	0·75	7·11342
14	1	8·76803	0·67	6·05710	0·75	5·91360
15	1	1·00092	0·67	6·25875	0·75	5·88210
16	1	1·60837	1	2·86575	0·75	7·22348
17	1	4·35200	1	7·94210	0·75	4·20166
18	1	3·62979	1	1·05613	0·75	1·18261
19	1	3·16162	1	8·61142	0·75	1·47634
20	1	3·44190	1	3·90078	0·75	4·61303

Table I. (Continued)

Correct model: $E(Y|x) = 5x$

(*c*) Regression analysis of Y_1 on X_1 (correct model)

Source	SS	d.f.	MS
Model	155·794907	1	155·794907
Residual	126·120725	18	7·00670695
Total	281·915632	19	14·8376649

Number of obs = 20
$F(1, 18) = 22·24$
Prob > $F = 0·0002$
R-squared = 0·5526
Adj R-squared = 0·5278
Root MSE = 2·647

y_1	Coefficient	Standard error	t	$P > \|t\|$	[95% confidence interval]	
x_1	5·582023	1·183783	4·715	0·000	3·094988	8·069058
_cons	− 0·048719	0·8370607	− 0·058	0·954	− 1·807318	1·70988

(*d*) Regression analysis of Y_2 on X_2 (correct model – Less efficient allocation of X's)

Source	SS	d.f.	MS
Model	24·6805628	1	24·6805628
Residual	78·5551996	18	4·36417775
Total	103·235762	19	5·43346117

Number of obs = 20
$F(1, 18) = 5·66$
Prob > $F = 0·0287$
R-squared = 0·2391
Adj R-squared = 0·1968
Root MSE = 2·0891

y_2	Coefficient	Standard error	t	$P > \|t\|$	[95% confidence interval]	
x_2	2·974769	1·250913	2·378	0·029	0·3466994	5·602839
_cons	0·9110929	0·7806436	1·167	0·258	− 0·7289785	2·551164

(*e*) Regression analysis of Y_3 on X_3 (correct model – design less efficient than given by X_1 [(*c*) analysis])

Source	SS	d.f.	MS
Model	78·713809	1	78·713809
Residual	100·825839	18	5·6014355
Total	179·539648	19	9·44945516

Number of obs = 20
$F(1, 18) = 14·05$
Prob > $F = 0·0015$
R-squared = 0·4384
Adj R-squared = 0·4072
Root MSE = 2·3667

y_3	Coefficient	Standard error	t	$P > \|t\|$	[95% confidence interval]	
x_3	5·290287	1·411248	3·749	0·001	2·325364	8·255209
_cons	0·546681	0·7484274	0·730	0·475	− 1·025707	2·119069

Table I. (Continued)

Correct model: $E(Y|x) = 11x - 6x^2$

(f) Regression analysis of W_1 on X_1 and X_1^2 (correct model, but design prevents estimating quadratic terms)

Source	SS	d.f.	MS
Model	64·9821823	1	64·9821823
Residual	79·6091957	18	4·42273309
Total	144·591378	19	7·61007252

Number of obs = 20
$F(1, 18) = 14·69$
Prob > $F = 0·0012$
R-squared = 0·4494
Adj R-squared = 0·4188
Root MSE = 2·103

w_1	Coefficient	Standard error	t	$P > \|t\|$	[95% confidence interval]	
x_1	3·605057	0·9405034	3·833	0·001	1·629133	5·580981
X_1^2	(dropped)					
_cons	0·634291	0·6650363	0·954	0·353	−0·7628985	2·03148

(g) Regression analysis of W_2 on X_2 and X_2^2 (correct model)

Source	SS	d.f.	MS
Model	91·3522269	2	45·6761134
Residual	126·624177	17	7·44848097
Total	217·976403	19	11·4724423

Number of obs = 20
$F(1, 17) = 6·13$
Prob > $F = 0·0099$
R-squared = 0·4191
Adj R-squared = 0·3508
Root MSE = 2·7292

w_2	Coefficient	Standard error	t	$P > \|t\|$	[95% confidence interval]	
x_2	13·30121	5·757084	2·310	0·034	1·154823	25·44759
X_2^2	−8·105359	5·520267	−1·468	0·160	−19·7521	3·541387
_cons	−0·0883785	1·188492	−0·074	0·942	−2·595878	2·419121

(h) Regression analysis of W_3 on X_3 and X_3^2 (correct model but design prevents estimating quadratic terms)

Source	SS	d.f.	MS
Model	127·606753	1	127·606753
Residual	70·9579032	18	3·94210573
Total	198·564656	19	10·4507714

Number of obs = 20
$F(1, 18) = 32·37$
Prob > $F = 0·0000$
R-squared = 0·6426
Adj R-squared = 0·6228
Root MSE = 1·9855

w_3	Coefficient	Standard error	t	$P > \|t\|$	[95% confidence interval]	
x_3	6·735821	1·183908	5·689	0·000	4·248523	9·223119
X_3^2	(dropped)					
_cons	−0·540429	0·6278619	−0·861	0·401	−1·859518	0·7786599

The regressions computed from the three designs are given in Tables I(*c*) to I(*h*). The first response model, $E(Y|x) = 5x$, is the correct model for the computer output in Tables I(*c*), (*d*) and (*e*). The second response model, $E(Y|x) = 11x - 6x^2$ is the correct model for the output in Tables I(*f*), (*g*) and (*h*). The output is taken from STATA.[2] Most standard statistical software packages produce similar output.

From Table I(*c*), the coefficient B is estimated as 5·58 (1·18) where (1·18) is the standard error. From I(*d*), the estimate of B is 2·97 (1·25) and from I(*e*), the estimate is 5·29 (1·41). These are close to the correct value of 5 (none is significantly different from 5). From Tables I(*f*) and I(*h*), where the correct model is $E(Y|x) = 11x - 6x^2$, we note that the quadratic coefficient C cannot be estimated since there are only two doses which are given to the subjects. To fit a quadratic model at least three distinct values of x are needed. The estimates fit a straight line between dose $x = 0$ and dose $x = 1$ (or $x = 0{\cdot}75$). Since $E(Y|x = 1) = 5$, and $E(Y|x = 0) = 0$, the slope is again 5, and the estimates 3·61 (0·94) and 6·74 (1·18) reflect that (from Tables I(*f*) and I(*h*)). The only design of the three which allows us to estimate C, (Table I(*g*)) gives 13·30 (5·76) as the estimate for B and $-8{\cdot}11$ (5·52) as the estimate for C. Neither estimate is significantly different from this parameter, which are $B = 11$ and $C = -6$, at the 0·05 significance level. The first and third designs do not permit estimation of some parameters. We note that the $N/2$ at 0 and $N/2$ at 1 is optimal for a simple linear regression with an intercept. It is not optimal for the model $E(Y|x) = 5x$ when the intercept is known to be 0. We see in these examples that the choice of design for a particular dose–response model is extremely important.

Determining patterns of dose responses is an important part of new drug evaluation. This may consist of a simple comparison of two levels (placebo and drug at some level), or may consist of placebo (dose $x = 0$) and multiple levels of the drug. We describe some of the design options the researcher has, and provide some guidance on choices. We consider first the equivalence of the two-group design and a two-dose design, and note some properties when the higher dose in the two-group design is less than the maximum dose in a dose–response design. We then review optimal designs and give the efficiencies of several candidate designs for the simple linear and quadratic regression models. For regression notation, we use capital letters (A, B, C) to denote the parameters, and lower case letters (a, b, c) to indicate their estimates. We assume normal errors with mean zero and variance σ^2 throughout. In all cases, we can use standard multiple regression software packages to estimate the parameters A, B, C and σ.

2. EQUIVALENCE OF TWO-GROUP DESIGN WITH DOSE–RESPONSE STUDIES

The two group design is equivalent to a linear regression with doses at two levels. The usual two-sample t-test for equality of means is equivalent to the test that the regression slope coefficient is zero. Assume that half of the observations are allocated to each dose (that is, equal sample sizes in each group). If the doses are coded 0 (placebo) and 1 (dose given) then the difference between treatments is $(\mu_1 - \mu_0)$, where μ_i is the mean response at the ith dose level. In the dose–response model, we assume the maximum dose given is coded as 1, with intermediate doses placed at values between 0 and 1. The usual analysis of this design is a regression (possibly polynomial). The two-group design is a special case of a dose response with only two levels. Denote the mean response in group i as $\bar{y}_i$. It is easy to show that the estimate of the slope is $b = (\bar{y}_1 - \bar{y}_0)$. This leads to the test of the coefficient being $\sqrt{N}(\bar{y}_1 - \bar{y}_0)/2s$, where N is the total sample size and s^2 is the regression mean squared error. This statistic is the two-sample t-test statistic. Simple modifications yield the unequal allocation case.

In the context of dose response the two-group design can be expanded. Thus far, we assumed that the maximum dose in the two-group design was 1. In such cases, the two-group design with

Table II. Sample size needed to detect $B/\sigma = 1$ ($\alpha = 0{\cdot}05$, $1 - \beta = 0{\cdot}9$)

K	2	3	4	5	6	7	8	9	10
N	43	64	76	85	91	95	99	101	104
n	22	22	19	17	16	14	13	12	11

n is the number needed per dose. Thus, $nK \geqslant N$
$B/\sigma = (\mu_1 - \mu_0)/\sigma$

half of the observations at each level is the optimum one for a linear regression $E(Y|x) = A + Bx$ (in the sense that var(b) is minimized). It may, however, be important to have fewer than half of the observations at the placebo or high dose levels (for example, for increasing the chance of receiving a hoped for effective treatment or for ethical reasons such as reducing the risk of side-effects), so the investigator might wish to place some observations at 0, some at 1, and the remainder at intermediate doses. The experimenter may even unbalance the design by making the number of observations at each dose unequal. Similarly, the 0 dose might be increased to x_0. This would not be a placebo controlled study, but would still be able to demonstrate effectiveness of x at $x = 1$ if there were an increasing response for increasing x. In the phase III drug approval context, regulatory agencies expect some of the dose to be at the level for which approval is sought. Thus, a dose–response study with doses at 0, 0·5 and 1·0 would not suffice for an approval at 0·75. Generally, dose–response studies are done in phase II. We will examine some examples of these when we consider the efficiencies of the designs in later sections.

Assume bivariate observations, (y, x), are taken, where x is the dose given and y is the response to the drug. The dose begins at 0 (placebo) and has a largest value of 1 (this can always be handled by appropriate scaling). With K equally spaced doses, we have values of x at 0, $1/(K-1)$, $2/(K-1), \ldots, (K-2)/(K-1)$, 1. If we assume a straight-line model, we have

$$E(Y|x) = A + Bx$$

where B is the amount of increase in $E(Y|x)$ for a one unit increase in the dose, x. That is, B is still $\mu_1 - \mu_0$, the mean difference between the drug at $x = 1$ and the placebo at $x = 0$. The variance of Y given x is σ^2, and $\text{var}(b) = \sigma^2/\Sigma(x_i - \bar{x})^2$.

With an equal number of observations and equal spacing between them, the denominator of var(b) is $\sigma^2 N(K+1)/[12(K-1)]$ where N is the total sample size and $N = nK$, and n is the sample size per group. This is another example of a uniform design. For testing H_0: $B = \mu_1 - \mu_0 = 0$ against a two-sided alternative H_1: $B \neq 0$, the sample size needed to detect a slope of B is given by

$$N = 12(K-1)(Z_{1-\alpha/2} + Z_{1-\beta})^2\sigma^2/[B^2(K+1)]$$

where α is the significance level, and β is the probability of a type II error and Z_x is the upper $100 \times$ th percentile of the standard normal distribution.

One can use the above formula to compute the needed sample size. For example, to detect $B/\sigma = 1$ ($(\mu_1 - \mu_0)/\sigma = 1$) with $\alpha = 0{\cdot}05$, and $1 - \beta = 0{\cdot}9$, the required sample sizes are given in Table II. The sample sizes are rounded to integer values. In practice the sample size N should be increased so they are divisible by K.

When $K > 2$, the multiple x values permit us to examine curvature or a polynomial response. The equally spaced doses are not necessarily the optimally spaced x values for such fits. In practice, if there is confidence that the model is not much wrigglier than a quadratic, the value of K should not be much higher than 3 or 4 because of the considerably larger sample size requirements.

Table III. Equivalent maximum two-group dose for K equally spaced doses

K	2	3	4	5	6	10
Dose	1·0	0·816	0·745	0·707	0·683	0·638

For the case with $K = 2$ and $x = 0$ and 1, the above produces the sample size for a two-group design with equal number observations in each group,

$$N = 4(Z_{1-\alpha/2} + Z_{1-\beta})^2 \sigma^2/(\mu_1 - \mu_0)^2$$

with the required sample size in each group as $N/2$. This is the usual formula for the two-sample problem.

Given the above, the investigator can consider several issues. First, if the investigator is certain that the response is linear and the maximum dose is the one which will be administered to patients, the optimal two-group design allocates half of the sample to the 0 (placebo) dose and half to 1 and no multiple dosage design ($K > 2$) is more efficient in the sense that no other design can have a smaller variance of b. However, the high dose may have some toxicity which the investigators want to avoid. The alternatives are to reduce the high dose to a lower dose at $x < 1$ or to conduct a study with multiple doses so that fewer patients receive the highest dose. If we compare a dose–response study that has equal allocation to equally spaced doses between 0 and 1 (a uniform design) to a two-group design at 0 and x, can the dose–response design have a smaller variance of b than the two-group design where the slope B is still $\mu_1 - \mu_0$ and the total sample size N does not exceed that of the two-group design? (Answer: Yes). Where does the dose–response design begin to be better? (Answer: it depends on x). How does the number of doses relate to the maximum level (x) in the two-group case? (Partial answer: if $x < 1/\sqrt{3}$ any multiple point design wins). How do the variances of b and c compare when the doses are evenly spaced versus optimal design placement? (Answer: see below). It is sometimes proposed to have fewer observations at dose = 0 and dose = 1 for ethical reasons (fewer patients at dose = 0 to have more patients receiving something, fewer at dose = 1 to have less potential toxicity). What is the effect of reducing the number of observations at the extreme doses on the variances of the parameter estimates? (Answer: the increase in variance can be pretty bad).

If the corresponding maximum dose given in the two-group design is $x = 0{\cdot}5$ (for example, 500 mg in a 1000 mg maximum dose study), the change in mean response would be half that of the maximum. The sample size required (Table II for $K = 2$) would be multiplied by 4, and 172 patients (86 per group) would be required. This is larger than a dose–response design with 10 levels of drug, and one would clearly prefer a dose–response design. If the maximum dose for two-group dose design is 0·75, then the number of patients required would be 76 (38 per group) and the two group design would require about as many subjects as a dose–response design with four dosages (at 0, 0·33, 0·67 and 1). A two-group study and a dose–response study will have the same sample size if the formulae for N are equal. Some algebra shows that a two-group study where the drug is given at dose x will have the same sample size requirement as a dose–response study with K doses equally spaced from 0 to 1 if $x^2 = (K + 1)/[3(K - 1)]$. This leads to Table III as a table of equivalent sample size studies.

This can be interpreted to mean that if the maximum dose in the two-group study is 0·745, a four level study with doses at 0, 0·33, 0·67 and 1 will provide estimate of B with the same precision. Assuming equal spacing, we can show that the equivalent dose is never less than 0·577 ($1/\sqrt{3}$). These results suggest that 3 to 5 levels in a dose–response study will provide most of the gain when the two group dose is less than the maximum dose in the dose–response study. When

the dose to be studied in the two-group study is the maximum dose that would be administered in a dose–response study, the dose–response study does not provide a gain in power or precision of estimate. If the maximum dose in a two-group study is less than 0·8, an increase in power can usually be realized with a dose–response study.

Dose–response (regression) designs enable us to compare the response to different drug levels and evaluate the responses for possible curvature. If we assume a quadratic response, $E(Y|x) = A + Bx + Cx^2$, we can find optimal designs which minimize var(c), var($y(x)$) for a given x, or the generalized variance (the determinant of the covariance matrix of the estimates). Here, $y(x)$ is the predicted value of y given x. These designs will provide estimates of the curvature (that is, the quadratic coefficient) and also allow us to estimate the linear dose–response. We next discuss these concepts.

3. CONSIDERATIONS IN CHOOSING DESIGNS

A major advantage that dose–response studies have over two-group studies is their ability to examine departures from linear response. For example, by plotting the responses against the doses we can examine (visually) if there are large departures from the assumed linear dose–response relationship. It is also possible to test formally this relationship using the pure error term when there are replicated observations at the different doses. This is fully explored in such texts as Rawlings[3] or Neter *et al.*[4] (In submissions to regulatory agencies, this should usually be indicated in the analysis plan submitted with the study proposal.)

Similarly, all of the powerful diagnostic tools of regression analysis (for example, residual analysis, influence statistics, normal probability plots) are available to the investigator (see Rawlings[3] or Neter *et al.*[4]). While one could argue that these are also available in the two-group design (it is a special case of the dose–response study with $K = 2$), residual analysis is not able to detect curvature in this case, and influence is the same for all observations when equal allocation is used.

If we assume that the model is quadratic,

$$E(Y|x) = A + Bx + Cx^2$$

the variances of the coefficients are found by inverting the $\mathbf{X}'\mathbf{X}$ matrix where $\mathbf{X}$ is the design matrix (see, for example, Neter *et al.*[4]). The first column of $\mathbf{X}$ is a column of 1s. The next columns are the values of x and x^2. To find the $\mathbf{X}'\mathbf{X}$ matrix for the quadratic regression model, we define

$$F = (\Sigma x_i)/N, \quad S = (\Sigma x_i^2)/N, \quad T = (\Sigma x_i^3)/N, \quad Q = (\Sigma x_i^4)/N$$

(standing for first, second, third and fourth (quartic) powers, respectively). The covariance matrix is

$$\Sigma = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2/[N(SQ + 2FTS - S^3 - T^2 - QF^2)] \begin{bmatrix} SQ - T^2 & ST - FQ & FT - S^2 \\ ST - FQ & Q - S^2 & SF - T \\ FT - S^2 & SF - T & S - F^2 \end{bmatrix}$$

giving

$$\text{var}(a) = \sigma^2[SQ - T^2]/[N(SQ + 2FTS - S^3 - T^2 - QF^2)]$$

$$\text{var}(b) = \sigma^2[Q - S^2]/[N(SQ + 2FTS - S^3 - T^2 - QF^2)]$$

$$\text{var}(c) = \sigma^2[S - F^2]/[N(SQ + 2FTS - S^3 - T^2 - QF^2)].$$

From the covariance matrix, the optimal design for estimating any one of the parameters can be found by minimizing the variance of its estimate. However, the optimization problem is usually a complicated one. For example, if interest is only in B, the problem becomes how to allocate observations in the closed interval $[0, 1]$ so that the quantity

$$(Q - S^2)/[N(SQ + 2FTS - S^3 - T^2 - QF^2)]$$

is minimized subject to the constraint that the number of observations at the x_i's sum to N. The solution to this and related problems is generally difficult. For this reason, we use an alternative approach using approximate optimal design theory.

4. A BRIEF REVIEW OF APPROXIMATE OPTIMAL DESIGN THEORY

An optimal design is one which minimizes (or maximizes) some function of the covariances of the parameter estimates. There are many optimal design criteria. Often they lead to complicated (or intractable) expressions to optimize. As a way of dealing with the problem of complicated algebraic expressions (like the one above), Kiefer[5] proposed the concept of approximate designs. Although there was controversy at the time, it is now an accepted way of solving a design problem. Generally, a design is defined by specifying the number of points (locations) in the interval where observations are taken, and the number (or proportion) of observations to be taken at each of these points. For example, if N is the total number of observations in the experiment, a design is defined by taking n_i observations at specified points x, $i = 1, 2, \ldots, k$ with $\sum n_i = N$. Alternatively, if ξ_i is the proportion of observations to be taken at x_i this is the same as taking $N\xi_i = n_i$ observations at x_i. The concept of approximate designs extends this notion of design by allowing the $N\xi_i$ to be non-integers. Consequently, approximate designs may require taking a fractional number of observations at a point. This is not a problem if N is large. In practice, we round to the nearest integer, subject to $\sum n_i = N$.

The main advantages of considering approximate designs are: (i) the optimization problem is simplified; (ii) frequently, the approximate optimal design is close to the optimal design.[5] Further background readings in this area are (in ascending order of difficulty) Atkinson and Donev,[6] Fedorov,[7] and Pazman.[8]

Define Ω as the space of the design points, x_i. In the context of this article, Ω is the closed interval $[0, 1]$. Let ξ be an approximate design (or simply a design from now on) giving probability mass ξ_i at the point x_i, $i = 1, 2, \ldots, K$. The analogue of the $\mathbf{X}'\mathbf{X}$ matrix for ξ is its information matrix defined by

$$\mathbf{M}(\xi) = \Sigma \mathbf{f}(\mathbf{x}_i)\mathbf{f}(\mathbf{x}_i)^t \xi_i$$

where x_i is the set of predictor variables. For example, consider the simple linear regression $\mathbf{f}(\mathbf{x})^t = (\mathbf{1}\ \mathbf{x})$. If ξ is a design with equal allocation at $x_1 = 0$ and $x_2 = 1$, we have $\xi_1 = 0{\cdot}5$ and $\xi_2 = 0{\cdot}5$, and

$$\mathbf{M}(\xi) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} 0{\cdot}5 + \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} 0{\cdot}5 = \begin{pmatrix} 1 & 0{\cdot}5 \\ 0{\cdot}5 & 0{\cdot}5 \end{pmatrix} \tag{1}$$

which is $1/N$ times the usual $\mathbf{X}'\mathbf{X}$ matrix.

Many practical optimality criteria are formulated in terms of $\mathbf{M}(\xi)$. As an illustration, Table IV lists some common criteria as convex functions, $\mathbf{H}(\mathbf{M}(\xi))$ of $\mathbf{M}(\xi)$ and reasons for using them. For ease of notation, we assume the underlying model for the first two criteria is quadratic, that is, $\mathbf{f}(\mathbf{x})^t = (\mathbf{1}\ \mathbf{x}\ \mathbf{x}^2)$. Generalizations to other models are straightforward. An optimal approximate

Table IV. Criteria for optimal designs

Optimality criterion	Interest	$H(\mathbf{M}(\xi))$
var(b)	Estimate B accurately	$\{\mathbf{M}(\xi)^{-1}\}_{2,2}$
var (c)	Estimate C accurately	$\{\mathbf{M}\{\xi)^{-1}\}_{3,3}$
L-optimality	Estimate response at $x = x_0$	$\mathbf{f}(\mathbf{x}_0)^t\mathbf{M}(\xi)^{-1}\mathbf{f}(\mathbf{x}_0)$
D-optimality	Estimate all parameters (A, B, C)	$-\log\lvert M(\xi)\rvert$
G-optimality	Minimize max var($\hat{y}(x)$)	$\max_{\mathbf{x}\in\Omega}\mathbf{f}(\mathbf{x})^t\mathbf{M}(\xi)^{-1}\mathbf{f}(\mathbf{x})$

The notation $\{\ \}_{2,2}$ and $\{\ \}_{3,3}$ refer to minimization of var(b) and var(c), respectrively, which are the second and third diagonal elements of $\mathbf{M}^{-1}$

design ξ° is one for which $\min H(\mathbf{M}(\xi)) = H(\mathbf{M}(\xi^\circ))$, where the minimization is taken over the set of all approximate designs on Ω.

The verification of an optimal approximate design is straightforward in many cases (Fedorov[7]). For example, if interest is in estimating all parameters, a D-optimal design is appropriate (see below). For linear models one can check if a given design, ξ, is D-optimal by verifying

$$\mathbf{f}(\mathbf{x})^t\mathbf{M}^{-1}(\xi)\mathbf{f}(\mathbf{x}) \leqslant p \quad \text{for all } \mathbf{x}\in\Omega \tag{2}$$

where p is the number of parameters in the model. For simple and quadratic linear regression models p is equal to 2 and 3, respectively. This condition is easily verified in practice. Corresponding checking conditions for the other criteria are available in Fedorov.[7] As an illustration how (2) might be used, return to the simple linear regression example with $H(\mathbf{M}(\xi)) = -\log(|\mathbf{M}(\xi)|)$, then if ξ° assigns equal numbers of observations at 0 and 1 the information matrix $\mathbf{M}(\xi^\circ)$ is as given in (1). After some algebra, the left hand side of (2) is seen to be $2(1-2x+2x^2)$. Since $p = 2$, the condition (2) is satisfied and ξ° is D-optimal. Thus, the equal allocation design is D-optimal. Other optimal designs can be similarly verified.

Under the assumption of normality of the errors, the D-optimality criterion seeks to minimize the volume of the confidence ellipsoid for the parameters. This is achieved by maximizing the determinant of the information matrix, $|\mathbf{M}^{-1}(\xi)|$ or minimizing the generalized variance, that is, $|\mathbf{M}(\xi)|$ over the set of all approximate designs. When interest is in only one parameter (or the response at a particular point), minimizing the variance of the estimator is reasonable. This is the rationale behind the first three criteria in Table IV. Minimizing the variance of b is the most relevant goal for simple linear regression. Minimizing the variance of c is important in quadratic regression. Minimizing the variance of the response at a point is a goal which is obviously met by placing all observations at the point, but this strategy would eliminate all possibility of estimating regression coefficients. G-optimality minimizes the maximum variance of a predicted value over the interval [0, 1]. Consequently, this criterion may be useful for estimating the response curve.

To evaluate the usefulness of a design, we use the idea of design efficiency. This is a number between 0 and 1 and has the interpretation that its reciprocal measures the number of times the design has to be replicated for it to do as well as the optimal design. The efficiency is the ratio of the criterion of the optimal design to the value of the criterion for the proposed design. For example, if we want to estimate the parameter B using design ξ, the efficiency of ξ is given by $\text{var}_{\text{opt}}(b)/\text{var}_{\xi}(b)$. For the quadratic model, the optimal design (ξ_3 in Table V) to minimize var(c) places $N/4$ points at 0, $N/4$ points at 1 and $N/2$ at 1/2. The D-optimal design (ξ_2 in Table V) places $N/3$ at each of these points.[5] Thus, for ξ_3 the variances of a, b and c are $4\sigma^2/N$, $72\sigma^2/N$, and $64\sigma^2/N$, respectively, and for ξ_2 the variances are $3\sigma^2/N$, $78\sigma^2/N$ and $72\sigma^2/N$. These are σ^2/N

times the diagonal elements of the inverse of the information matrices for ξ_2 and ξ_3. Therefore, the efficiency of the design ξ_2 for the estimation of C is $\text{var}_{\xi 3}(c)/\text{var}_{\xi 2}(c) = 64/72 = 0{\cdot}889$. These calculations are illustrated more fully in Section 5.

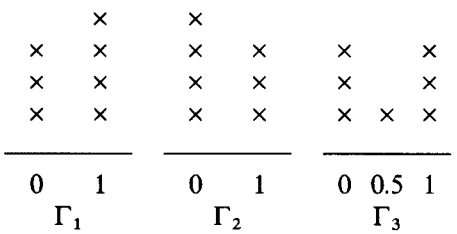
Sometimes this simple definition of efficiency must be modified to maintain its interpretation. For instance, consider D-optimality for a linear model with p parameters. If the D-optimal design is ξ_D, we have $|\mathbf{M}(\xi_D)| \geqslant |\mathbf{M}(\xi)|$ for all ξ, so that in order for ξ to do as well as ξ_D, ξ must be replicated, say r times. Since $|\mathbf{M}(\xi_D)| = |r\,\mathbf{M}(\xi)| = r^p|\mathbf{M}(\xi)|$, this implies $r = \{|\mathbf{M}(\xi)|/|\mathbf{M}(\xi_D)|^{-1/p}$ and so the D-efficiency of ξ is defined by $\{|\mathbf{M}(\xi)|/|\mathbf{M}(\xi_D)|^{1/p}$. For the two designs, ξ_2 and ξ_3, it can be verified that the D-efficiency of the design ξ_3 relative to ξ_2 is

$$\{|\mathbf{M}(\xi_3)|/\mathbf{M}(\xi_2)|\}^{1/3} = \{432/512\}^{1/3} = 0{\cdot}945$$

As a further example consider the case when $N = 2k + 1$, $\mathbf{f}(\mathbf{x})^t = (1\ \mathbf{x})$, $\Omega = [0, 1]$ and there is interest in three designs:

Γ_1 places $k/(2k + 1)$ of its mass at 0 and the rest at 1;
Γ_2 places $k/(2k + 1)$ of its mass at 1 and the rest at 0;
Γ_3 places $1/(2k + 1)$ of its mass at 1/2, $k/(2k + 1)$ at 0 and $k/(2k + 1)$ at 1.

With $N = 7$ ($k = 3$) graphically these three designs look like



It is easy to verify that $|\mathbf{M}(\Gamma_1)| = |\mathbf{M}(\Gamma_2)| = k(k + 1)/(2k + 1)^2$ and $|\mathbf{M}(\Gamma_3)| = k/[2(2k + 1)]$, so that $|\mathbf{M}(\Gamma_1)| - |\mathbf{M}(\Gamma_3)| = k/[2/(2k + 1)^2] > 0$ for all positive integers. As k becomes larger, the D-efficiency of Γ_3 approaches that of Γ_1 or Γ_2. For example, if $k = 3$, the efficiency is $[(2k + 1)/(2k + 2)]^{1/2} = 0{\cdot}9354$. For $k = 5$ ($N = 11$), the efficiency is 0·9574.

Kiefer[9] gave a general bound for the D-efficiency of a design which has a non-singular information matrix, $\mathbf{M}$. He showed that if $\mathbf{M}^{-1}(\Gamma)$ exists, then the D-efficiency is always greater than or equal to

$$\exp(1 - \max_{x \in \Omega}(\mathbf{f}(\mathbf{x})^t \mathbf{M}^{-1} \Gamma\, \mathbf{f}(\mathbf{x})/p).$$

Applying this to Γ_2 we have its D-efficiency is at least $\exp(-1/2k)$. For $k = 10$ ($N = 21$) this lower bound is 0·95. The D-efficiency approaches 1 quickly.

Our consideration of D-efficiency has been limited to the case where interest is in all the parameters. When nuisance parameters are present, techniques exist for estimating a subset of the model parameters. The analogous expressions for the checking condition (2) are invariably more complicated.[5]

5. CASE STUDIES FOR SIMPLE LINEAR AND QUADRATIC REGRESSION

We now study five dose–response designs and evaluate their efficiencies (see Table V). The dosage levels are 0, 1/3, 1/2, 2/3, 1 and the five designs offer different allocations to each. Because some dosages have no observations allocated to them, these are not five level designs in the usual sense.

Table V. Candidate designs for dose–response studies

Design	Dosage				
	0	1/3	1/2	2/3	1
ξ_1	$N/2$	0	0	0	$N/2$
ξ_2	$N/3$	0	$N/3$	0	$N/3$
ξ_3	$N/4$	0	$N/2$	0	$N/4$
ξ_4	$N/4$	$N/4$	0	$N/4$	$N/4$
ξ_5	$N/6$	$N/3$	0	$N/3$	$N/6$

Table VI. Design efficiencies for some designs for the linear and quadratic models on [0, 1]

Design	var (b) e_1	var(c) e_2	L-optimal var$(y\|x_0 = 1)$ e_3	D-optimal Generalized variance e_4	G-optimal max(var$(y(x))$ e_5
Simple linear regression: $y = A + Bx + \varepsilon$					
ξ_1	1·0000	NA	0·5000	1·0000	1·0000
ξ_2	0·6667	NA	0·4000	0·8165	0·8000
ξ_3	0·5000	NA	0·3333	0·7071	0·6667
ξ_4	0·5556	NA	0·3571	0·7454	0·7143
ξ_5	0·4074	NA	0·2895	0·6383	0·5790
Quadratic regression: $y = A + Bx + Cx^2 + \varepsilon$					
ξ_1	1·0000	0·0000	0·5000	0·0000	0·0000
ξ_2	0·0128	0·8889	0·3333	1·0000	1·0000
ξ_3	0·0139	1·0000	0·2500	0·9449	0·7500
ξ_4	0·0113	0·7901	0·2632	0·9048	0·7895
ξ_5	0·0099	0·7023	0·1833	0·7845	0·5500

As mentioned above, design ξ_1 is the optimal design for estimating B for the simple linear regression model, and is, in fact, optimal for other purposes as well (see Table VI). Design ξ_2 is optimal for jointly estimating all parameters in the quadratic model (that is, D-optimal). It is appealing also because it assigns an equal number of subjects to three dosages uniformly spaced between 0 and 1. Design ξ_3 is optimal for minimizing the variance of c in the quadratic model. Design ξ_4 has equal allocation to dosages uniformly spaced between 0 and 1, and is motivated by its simplicity and ease of explanation to clients. Designs ξ_2 and ξ_4 are special classes of uniform designs mentioned earlier. Besides being easy to construct, uniform designs also are robust when there is uncertainty in the regression model (Wiens[10]). Design ξ_5 is motivated by the ethical considerations of having fewer observations at the placebo and maximum doses.

Table VI shows the efficiencies of the five designs for the five optimality criteria. For any design, ξ, we will refer to these efficiencies as $e_1(\xi), e_2(\xi)$, etc. We note that none of these designs is L-optimal (that is, minimizes the variance of y at $x_0 = 1$). It is immediate that the L-optimal design places all observations at the point $x_0 = 1$. This design is useless for the other criteria, since it has zero efficiency for them.

Several interesting results are evident. If the simple linear regression model holds, design ξ_1 has design efficiencies of 1 except for $e_3(\xi_1)$. Thus, design ξ_1 is useful as it achieves several goals in the study, including, but not limited to, the first, fourth and fifth criteria. Unfortunately, this design is inefficient when the quadratic model holds. Its efficiencies are $e_2(\xi_1) = e_4(\xi_1) = e_5(\xi_1) = 0$ and

Table VII. Dosage levels for optimal estimation of polynomials (naive choices in parentheses)

Degree	X-values				
3	0	0·2764	0·7236	1	
	(0	0·3333	0·6667	1)	
4	0	0·1727	0·5000	0·8273	1
	(0	0·2500	0·5000	0·7500	1)
5	0	0·1175	0·3574	0·6426	0·8825	1
	(0	0·2000	0·4000	0·6000	0·8000	1)

$e_3(\xi_1) = 0{\cdot}5$. While design ξ_1 estimates B with the smallest variance (in the quadratic model) it estimates C and A with maximal variance (Preitschopf and Pukelsheim[11]). The practical implication is serious since design ξ_1 provides no information about C when the quadratic model holds. Interestingly, if we modify design ξ_1 to, say, ξ_1^*, so that half of the observations are taken at 0 and half at x $(0 < x \leqslant 1)$, then

$$e_1(\xi_1^*) = x^2$$

$$e_3(\xi_1^*) = x^2/(2x^2 - 4x + 4)$$

$$e_4(\xi_1^*) = x$$

$$e_5(\xi_1^*) = x^2/(x^2 - 2x + 2)$$

under the simple linear regression model.

The other entries in Table VI can be interpreted similarly. For the quadratic model, design ξ_2 gives a variance of c 1·125 (= 1/0·8889) times that of design ξ_3. This was the example noted earlier. This means that design ξ_2 requires 12·5 per cent more observations to attain the same variance for c as design ξ_3.

These computations demonstrate the importance of considering the efficiency of a design. A poor choice of a design wastes resources, while a carefully designed experiment can furnish more information with fewer resources. A poorly designed experiment could also, in an extreme case, produce little or no information after the experiment is run. Such would be the case if one wished to estimate C and design ξ_1 was used.

We have restricted attention to simple linear regression and quadratic regression models. Similar ideas apply to polynomial models of degrees higher than 2. For example, when a polynomial of degree 3, 4 or 5 is used to model the dose–response relationship and interest is in all the parameters, the approximate D-optimal designs take equal proportions at the dosage levels given in Table VII. The uniform design on $K + 1$ points are chosen as i/K, $i = 0, \ldots, K$. The optimal points are close to the naive choices and so the naive choices should pay only a slight penalty in the design criteria costs. Optimal designs for polynomials of higher degrees and optimal designs for estimating subsets of the parameters are available in Fedorov.[7]

Finally, we comment that if we consider approximate designs uniformly spaced on K points, there is no advantage in considering values of K greater than 4 for the simple linear regression model and K greater than 7 for the quadratic regression model (see Fedorov[7] for details). Table VIII shows that large values of K can result in substantial decreases in efficiency. For both these criteria, the efficiency decreases as K increases. Where possible one should avoid designs

Table VIII. Efficiency of a uniform design on K points for a quadratic model

Criterion	K				
	3	4	5	6	7
var(b)	0·8889	0·7901	0·7000	0·6372	0·5926
Generalized variance	0·9449	0·9048	0·8390	0·7946	0·7631

with excessive numbers of points. This is true both from the viewpoint of optimal design and from the logistics of conducting the experiment.

6. SUMMARY AND RECOMMENDATIONS

We have described dose–response designs and illustrated them with simple (artificial) examples. The dose–response design generally provides more information than the two-group design (with half of the points at 0 and half at 1) unless the response is linear. It places fewer points at the highest and lowest dose levels and provides information about non-linear response. There is only a small drop in efficiency in the designs as we move from the optimal, symmetric design to the uniform design. Designs which deliberately reduce the number of points at the extremes (dose = 0 and dose = 1) lose efficiency rapidly, and we do not generally recommend them. The approach and analysis we adopted here is based on optimal approximate design. The primary reason for doing this is that the design problem is simplified and the optimal approximate design that results provides a useful guide to the practitioner. There are many applications of the theory of optimal approximate design in practice and they are increasing. Some recent biomedical applications of these ideas can be found in Hoel and Jennrich,[12] Dunn,[13] Hatzis and Larntz,[14] Atkinson *et al.*,[15] and Kitsos *et al.*[16]

REFERENCES

1. Kotz, S. and Johnson N. L. (eds) *Encyclopedia of Statistical Sciences vol.* 2, Wiley, New York, 1982, p. 418.
2. StataCorp. *Stata Statistical Software: Release 4.0*, Stata Corporation, College Station, TX, 1995.
3. Rawlings, J. *Applied Regression Analysis: A Research Tool*, Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, CA, 1980.
4. Neter, J., Wasserman, W. and Kutner, M. *Applied Linear Statistical Models*, 3rd edn, Richard D. Irwin, Homewood, IL, 1990.
5. Kiefer, J. *Jack Carl Keifer Collected Papers III: Design of Experiments*, Springer-Verlag, New York, 1985.
6. Atkinson, A. C. and Donev, A. N. *Optimum Experimental Designs*, Clarendon Press, Oxford, 1992.
7. Fedorov, V. V. *Theory of Optimal Experiments*, translated and edited by Studden, W. J. and Klimko, E. M., Academic Press, New York, 1972.
8. Pazman, A. *Foundations of Optimum Experimental Design*, D. Reidel Publishing Company, Dordrecht, Boston, Lancaster and Tokyo, 1986.
9. Kiefer, J. 'Optimum experimental designs V, with application to systematic and rotatable designs', *Proc. 4th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, 1960, pp. 381–405.
10. Wiens, D. P. 'Designs for approximately linear regression: two optimality properties of uniform designs', *Statistics and Probability Letters*, **12**, 217–221 (1991).
11. Preitschopf, F. and Pukelsheim, F. 'Optimal designs for quadratic regression', *Journal of Statistical Planning and Inference*, **16**, 213–218 (1987).
12. Hoel, P. G. and Jennrich, R. I. 'Optimal designs for dose response experiments in cancer research', *Biometrika*, **66**, 307–316 (1979).

13. Dunn, G. 'Optimal designs for drug, neurotransmitter and hormone receptor assays', *Statistics in Medicine*, **7**, 805–815 (1988).
14. Hatzis, C. and Larntz, K. 'Optimal design in nonlinear multiresponse estimation: Poisson model for filter feeding', *Biometrics*, **48**, 657–671 (1992).
15. Atkinson, A. C., Chaloner, K., Hertzberg, A. and Juritz, J. 'Optimum experimental designs for properties of a compartmental model', *Biometrics*, **49**, 325–337 (1993).
16. Kitsos, C. P., Titterington, D. M. and Torsney, B. 'An optimum design problem in rhythmometry', *Biometrics*, **44**, 657–671 (1988).

3.2 Monitoring

TUTORIAL IN BIOSTATISTICS
BAYESIAN DATA MONITORING IN CLINICAL TRIALS

PETER M. FAYERS,[1*†] DEBORAH ASHBY[2] AND MAHESH K. B. PARMAR[1]

[1] *MRC Cancer Trials Office, 5 Shaftesbury Road, Cambridge CB2 2BW, U.K.*
[2] *Department of Mathematical Sciences, University of Liverpool, Liverpool L69 3BX, U.K.*

SUMMARY

Many clinical trials organizations use regular interim analyses to monitor the accruing results in large clinical trials. In disease areas such as cancer, where survival is usually a major outcome variable, ethical considerations may lead to a stipulated requirement for data monitoring of mortality. This monitoring has frequently taken the form of limiting interim analyses to be few in number, and specifying an extreme p-value of, for example, $p < 0{\cdot}001$ or $p < 0{\cdot}01$ as grounds for early termination of the trial. Group-sequential methods are also used. However, none of these approaches formally assesses the impact that the results of a clinical trial may have upon clinical practice. Thus a trial might be terminated early because of apparent treatment benefits, but might fail to influence sceptical clinicians to modify their future treatment policy. We discuss the application of Bayesian methods, including the use of uninformative, sceptical and enthusiastic priors, and demonstrate that the necessary calculations are both straightforward to perform and easy to interpret statistically and clinically. Methods are illustrated with interim analyses of a clinical trial in oesophageal cancer.

Statist. Med., **16**, 1413–1430 (1997)
No. of Figures: 0 No. of Tables: 1 No. of References: 20

1. INTRODUCTION

Interim analysis of accruing information in clinical trials is necessary in order to monitor for unexpectedly large treatment effects and for excess toxicity. In many clinical trials survival may be one of the main outcome measures, and it would clearly be unethical and unacceptable to continue recruiting patients to the trial if early results provide conclusive evidence of a convincing superiority of one or other treatment policies. These considerations have led many clinical trials organizations to institute formal procedures for regular monitoring and interim analyses of their trials, especially those trials which are large, have lengthy recruitment periods, and involve patient survival. In many cases the results of such monitoring are reviewed by specially convened Data Monitoring Committees.

* Correspondence to: P. M. Fayers, Unit for Epidemiology and Clinical Research, Faculty of Medicine, Medisinsk Teknisk Senter, N 7005 Trondheim, Norway

† Current address: Unit for Epidemiology and Clinical Research, Faculty of Medicine, Medisinsk Teknisk Senter, N 7005 Trondheim, Norway

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
 ISBN: 0-470-02365-1

Bayesian methods which may be used for data monitoring in clinical trials are illustrated, and a simple exposition provided showing how to apply these techniques. Previously published papers on this subject have tended either to be theoretical or to be wide ranging;[1-6] in all cases, details of the methods have been presented in a manner which may not be readily accessible to those who simply seek to apply Bayesian methods to their own trials but find the mathematics unfamiliar. In this paper the description of the methods should be sufficiently simple for non-mathematical readers to appreciate the objectives, to be able to perform the calculations, and to understand the interpretation of the results; however, mathematical details intended for applied statisticians are also included. Thus this tutorial adopts a position mid-way between exposition and 'cook-book', and should be suitable for both clinicians and statisticians involved in clinical trials.

Three worked examples are given, based upon data monitoring of the Medical Research Council (MRC) OE02 trial which is evaluating the role of surgery with or without adjuvant chemotherapy for treatment of patients with oesophageal carcinoma. Confidentiality precludes the publication of interim results, and so hypothetical scenarios are presented illustrating monitoring of: (i) a trial with positive results, but at an early stage of patient recruitment; (ii) the same trial at a later stage, when early termination would be recommended, and (iii) a trial which could be terminated because early results suggest there is unlikely to be any treatment difference. These examples relate to survival comparisons, which are very pertinent to trials in many disease areas, but adaptation to binary or continuous endpoints is relatively simple.

2. MONITORING OF CLINICAL TRIALS

When a trial accrues patients over several years, results on earlier patients become available before the later ones are randomized. If the results look promising in favour of one treatment, the question can arise as to whether the trial should be terminated early. In particular, if the early results provide reasonably conclusive evidence of an advantage in favour of one of the treatments, it may be considered unethical to continue recruiting patients. This especially applies to clinical trials in potentially fatal diseases, in which patients receiving an inferior therapy may be at higher risk of death. It is crucial that such studies should be closely monitored so that if the new treatment has an effect that is larger than expected the trial may be terminated early; it is equally crucial that if the new treatment is unexpectedly found to be inferior, the trial should also terminate early. However, even where survival is not the primary outcome, it may still be unethical to continue exposing patients to an inferior treatment. Furthermore, one can also argue that it is an abuse of research funds to continue even a harmless clinical trial beyond the point at which there is sufficient evidence as to which therapy more is effective. Thus many clinical trial protocols contain explicit statements about the frequency and timing of interim analyses, and many trials have a formal, independent Data Monitoring Committee which is assigned the task of overseeing the monitoring of the trial.[7]

Superficially, therefore, it might appear that clinical trials should be terminated as soon as there is a convincing and statistically significant difference between the treatments. Thus one might envisage a sequential or group sequential trial,[7,8] which would specify a stopping rule based upon formal statistical tests. This approach is being used in a few MRC trials, albeit with some reservations.[9] It is worth remembering that if a trial was important enough to start in the light of the knowledge available then caution should be exercised before too lightly concluding that there is sufficient evidence to terminate the trial; hence p-values alone are unlikely to suffice for decision making about the future of a clinical trial, although they may be an important consideration.

Nevertheless, there is an opposing school of thought which argues persuasively that the role of a clinical trial is to influence clinical opinion and clinical practice. Thus if a clinical trial detects a large treatment effect after half the patients have been entered, and as a consequence is terminated early, that trial may be received with considerable scepticism by clinicians; despite any significant p-values that are cited, many clinicians may still remain unconvinced by the weight of evidence that has been produced. These clinicians are likely to continue treating new patients in the same way that they have done in the past. The clinical trial, therefore, will have failed in its primary objective. Through early termination, it has failed to obtain sufficient evidence to alter the management and therapy of future patients. This philosophy has led many trialists to be cautious about stopping recruitment prematurely. The ISIS (International Study of Infarct Survival) trials, for example, explicitly state in the protocols that the Data Monitoring Committee will only disclose interim results to the steering committee if there is '**both** (a) "proof beyond all reasonable doubt" that for all, or for some, types of patient one particular treatment is clearly indicated or contraindicated in terms of net difference in mortality, **and** (b) evidence that might reasonably be expected to influence materially the patient management of many clinicians who are already aware of the other main trial results'; the protocol also suggests that this might perhaps correspond to a difference of at least three standard deviations.[10] The need to convince others is also formalized by the drug regulatory process, and a trial that has stopped early may fail to convince regulators; for this reason, too, many trialists are wary of premature termination of patient recruitment.

The concept of the role of a clinical trial being to influence clinical opinion has a number of important consequences. In particular, it implies that statistical significance and statistical stopping rules will not in themselves be sufficient, and that one should additionally consider the prior opinions of clinicians. If clinicians, in general, are sceptical about the merits of a new treatment in terms of its prolonging life or curing patients, the necessary evidence to change that view will have to be substantial; if, on the other hand, most clinicians already expect the new treatment to be an improvement, far less weight of evidence will be necessary to influence them.

In practice, most major trials groups use an independent Data Monitoring Committee to help with the review of trials. One of the functions of a Data Monitoring Committee is to offer advice on whether a particular trial should terminate, and although this is not only a statistical decision, statistical guidelines can help formalize and clarify some of the issues outlined above.[11] There are essentially two schools of thought concerning the statistical procedures and calculations that should be made. The first, adopting a 'classical frequentist' approach, pre-specifies a number of 'looks' at the accumulating data and uses the observed p-values of these looks as a basis for stopping. At each look a relatively stringent significance level is used, so that the overall level of significance for the trial is maintained at, say, 5 per cent. The Pocock rule,[8] for example, uses the same significance level at every analysis, whereas the O'Brien/Fleming rule[12] uses extremely stringent criteria at the very earliest visits on the grounds that early observed differences are much more likely to be spurious. These are examples of group-sequential designs. The second, or 'Bayesian', approach formalizes the idea that external or prior evidence or beliefs can be summarized mathematically, and that in stopping the trial one is balancing the evidence from the trial against this other evidence.[1–6,13] When the trial evidence can outweigh this other evidence it is time to stop the trial. Clearly the formalization of this other evidence is critical. This paper examines the practical aspects of specifying prior opinions and the application of a Bayesian approach.

3. DESIGN OF MRC OESOPHAGEAL TRIAL

As an example, we consider the MRC OE02 clinical trial. This aims to evaluate the role of pre-operative chemotherapy for patients with resectable cancer of the oesophagus.

The outlook for patients with oesophageal cancer undergoing surgery remains poor, with only 20 per cent remaining alive at 2 years, and only 5 per cent alive and disease-free at 5 years. However, results from several small, uncontrolled, phase II studies suggest that this cancer may respond to chemotherapy given either pre- or post-operatively. 2-year survival figures of as large as 30 to 40 per cent have been claimed. Two of the more active chemotherapy agents are cisplatin and fluorouracil. Hence OE02 is comparing survival for patients randomized to either pre-operative chemotherapy followed by surgery, or surgery alone. The chemotherapy in OE02 consists of two four-day courses of cisplatin and fluorouracil, with an interval of three weeks between courses. (Copies of the protocol may be obtained from the MRC Cancer Trials Office.)

However, the chemotherapy is expensive. It may sometimes have adverse side-effects including nausea and vomiting, and less frequently diarrhoea, stomatitis, renal disturbance and myelosuppression. Furthermore, since these patients will eventually undergo surgery, most of them would prefer the surgery to take place as soon as possible. Demonstrating equivalence of the two treatment arms is of no interest. Therefore OE02 is testing the hypothesis that pre-operative chemotherapy will improve survival, and that the patients' overall well-being is not impaired. Thus the primary endpoint of interest in OE02 is length of survival, although clearly the general well-being of the patients is also evaluated.

In accordance with MRC standard policy, a Data Monitoring Committee was created. This includes one independent statistician, and two independent clinicians who are experts in oesophageal cancer but are not entering patients into OE02. The trial was launched in 1992, and has a planned sample size of 800 patients. Over 400 patients were recruited by summer 1996. As this is an on-going clinical trial, the true interim results are confidential; the examples that follow are based upon fictitious data.

4. NOTATION

- $\sim N(\mu, \sigma^2)$ indicates 'is distributed as a normal (Gaussian) distribution with mean μ and standard deviation σ' (that is, a variance of σ^2)
- Φ is the cumulative normal probability, so that $\Phi(1{\cdot}6445) = 0{\cdot}95$.
- $\log(\ldots)$ represents logarithm to the base e.
- $\log(h)$ is the log-hazard ratio, and $\log(h_1)$ the log hazard ratio under the alternative hypothesis, H_1.
- We assume a clinical trial is being carried out, and that at the time of carrying out the interim analysis we have observed O_1 and O_2 deaths or 'events' in the two treatment groups.
- N_d, the total number of deaths observed to date, is given by

$$N_d = O_1 + O_2. \tag{1}$$

- E_1 and E_2 are the 'expected' number of deaths that would have been observed under the null hypothesis; computer programs which calculate survival comparisons usually display E_1 and E_2.

5. HAZARD RATIOS AND SURVIVAL TRIALS

Suppose the clinical trial was designed to compare survival in patients randomized between a standard form of treatment versus a new treatment. Frequently there will be prior knowledge

about the nature of the survival curve for the standard treatment. This may be derived from previous studies or from clinical experience. For example, in the OE02 trial past experience enabled us to expect that 20 per cent of patients receiving standard surgical treatment would still be alive at 2 years after surgery. Thus the 2-year survival rate is 0·20. More generally, we use $surv_1$ and $surv_2$ to represent the survival rates for the two treatment groups, where survival is measured at some fixed time relative to randomization. Thus $surv_1$ might be the pre-study estimate of survival in the standard or control arm of the trial, and would represent the proportion of patients expected to be alive at some specified time point relative to when the patient was randomized. $Surv_2$ would be the survival rate that is hypothesized for the alternative treatment. The trial will have been designed to test a null hypothesis of no treatment difference, against an alternative hypothesis that the treatment difference is at least $surv_2$-$surv_1$.

An estimate of $surv_1$, together with a target value for the alternative hypothesis of a treatment difference of at least $surv_2$-$surv_1$, is usually specified in clinical trial protocols and is used as a basis for sample size estimation (see example 1(a)). The sample size calculations ensure that when the clinical trial has been completed, and provided there has been adequate follow-up of the patients, the trial-based estimates of $surv_1$ and $surv_2$ will be sufficiently precise to enable adequately powerful hypothesis testing; a review of sample size issues is given in Fayers and Machin,[14] and tables for sample size estimation are available.[15,16]

In terms of hazard ratios, this is equivalent to carrying out a trial to detect a log hazard ratio, which we call $\log(h_1)$, of

$$\log(h_1) = \log(\log(surv_1)/\log(surv_2)). \tag{2}$$

Hence we have a null hypothesis that the log hazard ratio is zero, and an alternative hypothesis that the log hazard ratio is $\log(h_1)$.

5.1. Example 1(a)

In OE02 the baseline proportion surviving 2 years, in patients receiving surgery alone, was assumed to be 0·20 (20 per cent of patients remaining alive after 2 years). The alternative hypothesis is that pre-operative chemotherapy produces an absolute improvement of 10 per cent, to 0·30 at 2 years. These values were used as a basis for the sample size estimation, and are specified in the study protocol. From equation (2), this translates to an alternative hypothesis with a log hazard ratio of $\log(h_1) = 0{\cdot}290$.

6. INTERIM ANALYSES

When a trial is being monitored there will be interim, and less precise, estimates of $surv_1$ and $surv_2$ which are based upon the survival experience of patients currently recruited to the trial and followed up until the time of the analysis. These estimates allow calculation of the data-based log hazard ratio $\log(h_d)$, as will be shown later. The role of the interim analysis, and indeed the application of methods described in this paper, is to assess the currently available information from the trial so as to determine whether there is already sufficiently convincing accruing evidence from $\log(h_d)$ for the Data Monitoring Committee to consider terminating patient recruitment to the trial.

7. THE BAYESIAN APPROACH: PRIORS, LIKELIHOODS AND POSTERIORS

The Bayesian approach formalizes the procedure of having pre-study beliefs, which are then influenced by the results from an experiment such as a clinical trial, to yield revised beliefs. These pre-study beliefs are expressed as a 'prior distribution'. If we consider a therapy trial such as OE02, a clinician may start by believing, for example, that the absolute treatment difference is likely to be 10 per cent. Perhaps it is thought possible, although less likely, that the difference could be as large as 15 per cent or as small as 5 per cent; furthermore, it might be thought just about possible, but very unlikely, that it could range from 20 per cent down to 0 per cent. Some clinicians might even fear that the chemotherapy, by delaying surgery, could conceivably confer a survival disadvantage. By persuading the clinician to quantify the terms 'is likely', 'less likely', and 'very unlikely' in terms of probabilities (for example, some clinicians might ascribe the words 'less than 5 per cent of the time' to 'very unlikely'), we can build up a probability distribution for the chance of the treatment effect being of various magnitudes. This is called the 'prior distribution'.

After carrying out the study, we can evaluate the chance that we would have obtained such extreme data if the effect size were of a specified magnitude. This probability is called the 'data likelihood', and can be evaluated across the range of all plausible values for the effect size.

Finally, we have the 'posterior distribution', which is simply the prior distribution modified as a consequence of the observed results. This posterior distribution provides an estimate of what we would expect that same clinician to believe if realistic allowance is made for the information obtained from the trial.

Bayesian methods are becoming increasingly widely employed in clinical trials, although many of these applications remain controversial; a special issue of *Statistics in Medicine*[17] reviewed the state of art. For many statisticians the principal reason for disquiet with Bayesian techniques is the determination of suitable prior distributions. However, in the context of *monitoring* of trials, this becomes a largely specious issue, as we shall show. Thus even a traditional statistician, following the classical frequentist approach, is likely to find Bayesian monitoring not merely satisfactory but more appropriate than alternative approaches.[1,2,13]

8. BAYES' THEOREM

Bayes' theorem puts into mathematical notation the concept of having a prior belief which is then modified according to the observed data, resulting in a revised posterior belief. Formally, it states that the posterior distribution after observing data is proportional to the data 'likelihood' multiplied by the prior distribution. Thus if H is a hypothesis concerning a parameter, such as the treatment effect size (for example, the log hazard ratio of a survival comparison), we have:

$p(H)$ equals the pre-study opinion (prior probability) about the treatment effect size, and

$p(\text{data}|H)$ equals the likelihood of obtaining the observed data, given the effect size.

Then $p(H|\text{data})$ is the revised opinion (posterior probability) about the treatment effect size, given the observed results. This is proportional to the product of the prior and the data-likelihood, which can be written as:

$$p(H|\text{data}) \propto p(\text{data}|H)\, p(H). \tag{3}$$

This is the fundamental equation that underpins the Bayesian approach, and which we shall later apply to the prior distributions which represent clinicians' opinions.

9. PRIOR DISTRIBUTIONS

A prior distribution is chosen to provide an estimate of initial beliefs concerning the size of the potential therapeutic benefit. Thus, if one hopes that the results from a clinical trial will influence the treatment of future patients, the prior distribution should represent the level of scepticism that is expressed by those clinicians that one seeks to influence.

There are three principal types of prior we could use, which are the uninformative, or reference, prior, sceptical prior and enthusiastic prior.[1,2] Other priors that could be considered are a clinical prior (often similar to the enthusiastic prior) and a meta-analysis prior;[1-3] this paper only discusses the first three.

9.1. Uninformative prior

The uninformative, or reference, prior represents a lack of clinical opinion as to the likely treatment difference, and in that sense contains no information about prior beliefs or other prior knowledge; hence the name 'uninformative prior'. It corresponds more or less to the conventional 'frequentist' approach of significance testing. One possible and adequate approximation is to assume that this prior can be represented by a normal distribution with mean 0 (corresponding to the null hypothesis of no difference) and an infinite variance. Although this is an 'improper' distribution (since a normal distribution cannot really have an infinite variance), it serves as a convenient mathematical device which in practice yields sensible posterior distributions. For analogy with the other prior distributions being considered, we will take this infinite variance to be 4/0:

$$\text{Uninformative prior} \sim N(0, 4/0).$$

9.2. Sceptical prior

The sceptical prior is specified by considering there is only a small probability that the alternative hypothesis (or better than it) is likely to be true. Thus a sceptical prior distribution may be given by considering the best guess to be zero, but allowing that there is a small probability, say γ, of an effect as large as or larger than $\log(h_1)$.

Interestingly, it can be shown that if we take $\gamma = 5$ per cent, and we carry out five interim analyses of the data, then the overall type I error is about 5 per cent (assuming the trial was designed with 90 per cent power, 5 per cent significance level, and a particular stopping rule).[18] Furthermore, if we compare this approach with classical methods, adopting a sceptical prior gives us a procedure which lies between the Pocock rule and the O'Brien/Fleming rule.[2] It should be noted that the role of the sceptical prior is to guard against over-enthusiastic acceptance of a positive and possibly large treatment effect that might be observed by chance at the time of an interim analysis. In this situation emphasis on probabilities in only one direction is appropriate, and this prior is not appropriate if the treatment effect appears to be in the opposite direction. Thus a one-sided probability is used.

Setting $\gamma = 5$ per cent, we can solve for σ_{scep}:

$$1 - \Phi\left(\frac{\log(h_1)}{\sigma_{\text{scep}}}\right) = \gamma = 0{\cdot}05$$

$$\frac{\log(h_1)}{\sigma_{\text{scep}}} = 1{\cdot}6445$$

$$\sigma_{\text{scep}} = \log(h_1)/1{\cdot}6445. \tag{4}$$

Although a normal distribution may not be entirely correct for a sceptical prior, it provides a convenient assumption which has been supported by empirical evidence; studies have been carried out, showing that clinical opinions commonly result in a log hazard ratio with a normal prior distribution.[2]

Therefore we adopt a sceptical prior which is represented by a normal distribution with zero mean and variance σ^2_{scep}. Hence we have the sceptical prior $\sim N(0, \sigma^2_{\text{scep}})$.

Now for survival data the variance of the log hazard ratio is approximately $4/n$ where n is the total number of events; this is a remarkably good approximation.[19] This approximation enables us to determine an equivalent study size corresponding to this sceptical prior. Equating the sceptical prior variance to $4/n$, we have $\sigma^2_{\text{scep}} = 4/n$. This can be solved for n, and thus the sceptical prior is equivalent to having performed a trial with $N_p = n$ patients, all of whom have died, and in which no difference has been observed between the two arms. Hence:

$$N_p = 4/\sigma^2_{\text{scep}} \tag{5}$$

giving the sceptical prior $\sim N(0, 4/N_p)$.

9.3. Example 1(b)

For the MRC OE02 clinical trial, we obtained $\log(h_1) = 0{\cdot}290$. Solving equation (4) gives $\sigma_{\text{scep}} = 0{\cdot}176$, and from equation (5) we see that this is equivalent to having performed a trial with $N_p = 129$ patients, all followed to death.

9.4. Enthusiastic prior

An enthusiast might argue that the treatment difference is bound to be greater than zero and that, as a best guess, it is likely to be equal to $surv_2$-$surv_1$. Hence an enthusiastic prior can be taken by considering the best guess to be the alternative hypothesis. We also assume this has the same precision as the sceptical prior; therefore we assume a normal distribution with mean $\log(h_1)$ and variance $4/N_p$:

$$\text{Enthusiastic prior} \sim N(\log(h_1), 4/N_p).$$

10. JUSTIFICATION FOR OUR CHOICE OF PRIOR DISTRIBUTIONS

We have described three prior distributions, although there are many potential distributions that could be considered; Parmar *et al.*[2] and Spiegelhalter *et al.*[1] discuss a few other possibilities. The difficulty of making an objective and non-controversial selection is, of course, one of the reasons why frequentist statisticians often object to Bayesian techniques. Thus, for example, if the outcome of a clinical trial is being reported using a Bayesian approach, there is always the anxiety that the investigators may have chosen an unduly optimistic prior distribution, and thus may have unfairly claimed that they have confirmation of a treatment effect.

In the context of monitoring of clinical trials, many of these issues which are sometimes levelled against Bayesian methods are largely irrelevant. In particular, for monitoring, one attempts to

guard against premature termination of a clinical trial and thus chooses a prior distribution which reduces the chance of wrongly claiming that the results have already become conclusive. In effect, the role of Bayesian monitoring is to determine whether the early results are already so overwhelmingly convincing that there is no need to continue to collect further confirmatory information.

Hence, in general, the sceptical prior distribution is appropriate for monitoring positive results, when interim analyses appear to be indicating a substantial treatment effect that might lead to stopping the trial. On the other hand, the enthusiastic prior distribution is of greatest importance when the results are leaning towards equivalence, and as a brake against early termination of a trial when initial results suggest a detrimental apparent effect of the new treatment. Choosing these two priors provides a useful brake against premature termination of trials.

It will often be useful to present the results of analyses under several alternative prior distributions, so that the impact of the observed data may be reviewed according to different levels of scepticism. The uninformative prior has the advantage of corresponding roughly to the classical frequentist approach, and so this together with the sceptical and enthusiastic priors gives a broad and useful overview of the implications of terminating a clinical trial.

11. DATA

Having chosen one or more prior distributions, we then consider the observed data.

We can estimate h_d from the observed and expected deaths, using

$$\log(h_d) = \log\left(\frac{(O_1/E_1)}{(O_2/E_2)}\right). \tag{6}$$

We again assume that the log hazard ratio follows a normal distribution, so that the observed data have a normal distribution with mean h_d and variance $4/N_d$. Hence

$$\text{data} \sim \text{N}(\log(h_d), 4/N_d).$$

12. POSTERIOR DISTRIBUTIONS

The prior distribution may be based upon initial or prior beliefs and prejudices, and external evidence such as reports of other (possibly small, possibly non-randomized) studies. When data, in the form of deaths, accrue from the clinical trial, the prior beliefs should be modified according to the weight of evidence that has been collected. Statistically, we use the distribution of the observed data to modify our prior distributions, resulting in a 'posterior' distribution which reflects our revised beliefs in the light of the new data.

12.1. Statistical derivation of posterior distributions

The standard Bayes equation (3) allows a prior distribution to be modified to reflect the data that have been observed, yielding the posterior distribution. Specifically, we can adapt (3) and apply it to the distributions discussed above.

The three prior distributions may be written in the general form of $\text{N}(\mu_x, 4/N_x)$ where μ_x and N_x are 0,0 (uninformative prior), $0, N_p$ (sceptical prior), or $\log(h_1), N_p$ (enthusiastic prior).

Thus, corresponding to $p(H)$ in equation (3), we have

$$\text{Prior} \sim \mathrm{N}(\mu_x, 4/N_x).$$

Similarly, corresponding to $p(\text{data}/H)$, writing $\mu_d = \log(h_d)$, we have

$$\text{Data likelihood} \sim N(\mu_d, 4/N_d).$$

Then by $p(H|\text{data}) \propto p(\text{data}|H)\,p(H)$ we obtain the posterior distribution from the product of the prior and the likelihood. It can be shown that solving the equations gives

$$\text{Posterior} \sim \mathrm{N}\left(\frac{N_x\mu_x + N_d\mu_d}{N_x + N_d}, \frac{4}{N_x + N_d}\right).$$

By applying these equations we can derive the posterior distributions which reflect the impact of the currently observed data upon the uninformative, sceptical and enthusiastic priors. The full set of distributions are listed below.

13. SUMMARY OF EQUATIONS:

We consider the following prior distributions:

$$\text{Uninformative prior} \sim \mathrm{N}(0, 4/0)$$

$$\text{Sceptical prior} \sim \mathrm{N}(0, 4/N_p)$$

$$\text{Enthusiastic prior} \sim \mathrm{N}(\log(h_1), 4/N_p).$$

For the observed data, we have

$$\text{Data 'likelihood'} \sim \mathrm{N}(\log(h_d), 4/N_d).$$

This gives posterior distributions:

$$\text{Uninformative} \sim \mathrm{N}(\log(h_d), 4/N_d) \tag{P1}$$

$$\text{Sceptical} \sim \mathrm{N}\left(\frac{N_d\log(h_d)}{N_p + N_d}, 4/(N_p + N_d)\right) \tag{P2}$$

$$\text{Enthusiastic} \sim \mathrm{N}\left(\frac{N_p\log(h_1) + N_d\log(h_d)}{N_p + N_d}, 4/(N_p + N_d)\right) \tag{P3}$$

14. EVALUATION OF POSTERIOR DISTRIBUTIONS/REVISED BELIEFS

The equations (P1), (P2) and (P3) represent the distributions of the revised beliefs, based upon the pre-study prior beliefs which have been modified in the light of the current data. Now that the equations for these posterior distributions have been derived, we can calculate the probabilities of certain levels of improvement for each of the corresponding prior distributions and the given data.

Let us suppose that the target improvement is δ. From (2), we have

$$\log(h_\delta) = \log(\log(surv_1)/\log(surv_1 + \delta)). \tag{7}$$

This can be substituted into equations (P1), (P2), (P3), to produce a table, as in the following example.

14.1. Example 1(c)

Equations (P1), (P2) and (P3) can be used to construct Table I, showing 0 per cent, 5 per cent and 10 per cent absolute improvement in percentage survival. The reason for selecting these values is that 0 per cent corresponds to the null hypothesis, 10 per cent is the alternative hypothesis for the example trial, and 5 per cent is both mid-way between 0 per cent and 10 per cent and arguably a realistic yet still worthwhile survival benefit; if the survival gain were as low as 1 per cent, it is unlikely that clinicians would use the toxic and expensive chemotherapy, but if it were 5 per cent it might be a sufficiently large improvement to warrant recommending adopting pre-operative chemotherapy as the treatment of choice. Different percentages will be appropriate for other trials. Note that A is the probability that the improvement is greater than 0 per cent when a uniform prior is assumed, and that 1 − A corresponds roughly to a conventional p-value (one-sided) where a single test is made (that is, without allowance for multiple looks at the data). Similarly, 1 − B can be shown to be equivalent to a significance level corresponding to a monitoring rule lying between the Pocock and the O'Brien/Fleming rules.[2]

Table I

Target improvement δ	Log hazard ratio $\log(h_\delta)$	Probability that improvement is greater than the target value		
		Uninformative prior (P1)	Sceptical prior (P2)	Enthusiastic prior (P3)
0%		A	B	
5%			C	
10%				

15. STOPPING CRITERIA

The table showing the probabilities associated with various target improvements can be reviewed by the Data Monitoring Committee. However, although the Bayesian framework is useful for interpreting the current trial's results in an informal manner, it is still useful to have guidelines for when seriously to consider stopping. If the trial is between two treatment arms, then a reasonable criterion is to demand that the posterior probability of one treatment being better, in the light of a sceptical prior belief, is at least 95 per cent; thus cell B should exceed 95 per cent. Alternatively, if a non-zero target improvement is sought, a reasonable criterion might be to accept a posterior probability of 90 per cent; in Table I, for a 5 per cent improvement this would imply that cell C should be at least 90 per cent.

16. CALCULATIONS: SUMMARY

We start with a table giving the observed and expected number of events in each group, as is readily obtained from most survival analysis programs.

$N_d = O_1 + O_2$ = the total number of deaths observed to data.

Log(h_d), the log hazard ratio, can be obtained from equation (6)

$$\log(h_d) = \log\left(\frac{(O_1/E_1)}{(O_2/E_2)}\right)$$

Equations (2), (4) and (5) enable N_p to be estimated for the sceptical prior:

$$\log(h_1) = \log(\log(surv_1)/\log(surv_2))$$

where $surv_1$ and $surv_2$ are the hypothesised event rates.

$$\sigma_{\text{scep}} = \log(h_1)/1{\cdot}6445$$

$$N_p = 4/\sigma^2_{\text{scep}}.$$

Then the list of summary equations can be expanded out, although only the posterior distributions (P1), (P2) and (P3) are used for constructing the table of probabilities of various improvements, δ.

Finally, equation (7) allows $\log(h_\delta)$ to be evaluated for the improvements δ; this value of $\log(h_\delta)$ can be substituted for $\log(h_d)$ in (P1), (P2) and (P3), giving the required table of probabilities.

17. EXAMPLES

Three worked examples are given, illustrating monitoring of clinical trials such as the oesophageal trial where survival is the main endpoint. These illustrate the calculations for three different arios, and indicate the suggested interpretation of the results. In particular, example 1 (which is a continuation of the example already used in the text) is a trial which might be regarded as beginning to show emerging treatment differences, but is at an early stage of patient recruitment; this trial should be continued. Example 2 is the same trial at a later stage, when early termination would be recommended, and example 3 is a trial which could be terminated because early results suggest there is unlikely to be any treatment difference.

17.1 Worked example 1

The following is an example of what might happen if early results for the OE02 trial were to suggest that there might be a treatment effect, and the Data Monitoring Committee wished to consider the possibility of early termination. We shall adopt the position of a sceptic and mainly focus upon the sceptical prior.

We assume that routine interim analyses are being carried out, say annually. At the time of the first analysis perhaps 200 patients have been entered into the trial, of whom 100 have died (60 in group 1, and 40 in group 2), out of those so far recruited and entered into the trial. Standard survival-analyses techniques allow computation of the expected number of deaths in each group; these values are calculated and printed out by most survival analysis software.

Suppose the following results were obtained at the first interim analysis of the MRC OE02 trial:

Data:

Group	Observed	Expected	Obs/Exp
1	60	48·00	1·250
2	40	52·00	0·769

Applying the equations in Section 16, we obtain the log hazard ratio:

$$\text{Hazard ratio} = 1{\cdot}25/0{\cdot}769 = 1{\cdot}625$$

$$\text{Log hazard ratio} = 0{\cdot}486.$$

From parts 1(a) and 1(b) of this example, already given in the text, we have $surv_1 = 0{\cdot}20$ and $surv_2 = 0{\cdot}30$, giving

$$\log(h_1) = \log(\log(0{\cdot}20)/\log(0{\cdot}30)) = 0{\cdot}290.$$

Hence $\sigma_{\text{scep}} = \log(h_1)/1{\cdot}6445 = 0{\cdot}176$, and $N_p = 4/0{\cdot}176^2 = 129$ patients.

Priors: From the summary of equations, we have the following priors, data likelihood and posteriors:

$$\text{Uninformative prior} = \text{N}(0, 4/0)$$

$$\text{Sceptical prior} = \text{N}(0, 4/129)$$

$$\text{Enthusiastic prior} = \text{N}(0{\cdot}290, 4/129)$$

$$\text{Data likelihood} = \text{N}(0{\cdot}486, 4/100)$$

$$\text{Uninformative posterior} = \text{N}(0{\cdot}486, 4/100) \quad \text{from (P1)}$$

$$\text{Sceptical posterior} = \text{N}(0{\cdot}212, 4/229) \quad \text{from (P2)}$$

$$\text{Enthusiastic posterior} = \text{N}(0{\cdot}376, 4/229) \quad \text{from (P3).}$$

The table showing the probabilities of various improvements can be constructed by estimating the log hazard ratio for the target improvement, and using the three posterior distributions.

For example, for a target improvement of 0·100 upon the assumed baseline survival rate of 0·20 in OE02, we use equation (7) to obtain an equivalent log hazard of

$\log(\log(0{\cdot}20)/\log(0{\cdot}20 + 0{\cdot}100)) = 0{\cdot}290$. We then apply the enthusiastic posterior of N(0·376, 4/299) giving $1 - N\{(0{\cdot}290 - 0{\cdot}376)/\sqrt{(4/229)}\} = 0{\cdot}741$, as in the bottom right-hand cell.

Probabilities of improvement being greater than:

Target improvement	Log hazard	Uninformative	Sceptical	Enthusiastic
0·000	(0·000)	0·992	0·946	0·998
0·050	(0·149)	0·954	0·683	0·957
0·100	(0·290)	0·836	0·277	0·741

Interpretation: Our interpretation is that, starting from a sceptical position, there is reasonable (94·6 per cent) evidence of some improvement, but only modest evidence (68 per cent) of a difference as large as a 5 per cent improvement in survival, and only slim evidence (28 per cent) of a difference as big as 10 per cent.

Recommendations: Casting a sceptical eye over the data available, we would recommend that:

(i) the trial should continue at present;
(ii) the Data Monitoring Committee should consider recommending termination of the trial if the sceptical posterior probability of a 5 per cent difference exceeds 90 per cent.

17.2. Worked example 2

Suppose the study in example 1 continued recruiting patients, with the same observed hazard ratio, until 350 patients had entered and 200 deaths had occurred. The Data Monitoring Committee might now consider that there is increasing weight of evidence in favour of a treatment effect, and would again wish to re-examine the impact of the data upon someone adopting a sceptical stance.

Repeating all the calculations would yield the following:

Design: P1 = 0·20, P2 = 0·30, log hazard = 0·290, gamma = 0·05, SD scep = 0·176 (equivalent to study of 129 patients).

Data:

Group	Observed	Expected	Obs/Exp
1	120	96·00	1·250
2	80	104·00	0·769

Hazard ratio = 1·625

Log hazard ratio = 0·486.

Priors:

Uninformative prior = N(0, 4/0)

Sceptical prior = N(0, 4/129)

Enthusiastic prior = N(0·290, 4/129)

Data likelihood = N(0·486, 4/200)

Uninformative posterior = N(0·486, 4/200)

Sceptical posterior = N(0·295, 4/329)

Enthusiastic posterior = N(0·409, 4/329).

Probabilities of improvement being greater than:

Target improvement	Log hazard	Uninformative	Sceptical	Enthusiastic
0·000	(0·000)	1·000	0·996	1·000
0·050	(0·149)	0·991	0·907	0·991
0·100	(0·290)	0·916	0·518	0·859

Interpretation: We now see that a person who demanded an absolute change of 5 per cent in survival as the minimal worthwhile improvement, and who had a sceptical prior belief, would be likely to be convinced by the observed data, having a posterior probability of 90·7 per cent.

Recommendations: Casting a sceptical eye over the data available, we would recommend that the trial should cease recruitment of patients.

17.3. Worked example 3

Another situation might arise in which early results make it appear as if there is no difference between the treatments. For example, one might have entered about half the patients and might have observed death rates which suggest that the two treatments are either equal or that the new treatment is inferior. Would it be worth continuing the trial? In the case of the OE02 trial, one should remember that we are seeking to establish whether there is a treatment advantage (equivalence is not of interest). Thus we might decide that we would consider early termination if even an enthusiast would agree that it is very unlikely that the new treatment could provide clinically worthwhile treatment benefit.

Suppose that the following situation were to occur during the monitoring of the clinical trial. In this case we have recruited 450 patients and observed 300 deaths, and these are divided almost equally between the two treatment groups. Is it worth continuing the trial?

Here, again, we assume that a 5 per cent difference in survival would be considered worthwhile.

Data:

Group	Observed	Expected	Obs/Exp
1	153	150·00	1·020
2	147	150·00	0·980

$$\text{Hazard ratio} = 1{\cdot}041$$

$$\text{Log hazard ratio} = 0{\cdot}040.$$

Priors:

$$\text{Uninformative prior} = N(0, 4/0)$$

$$\text{Sceptical prior} = N(0, 4/129)$$

$$\text{Enthusiastic prior} = N(0{\cdot}290, 4/129)$$

$$\text{Data likelihood} = N(0{\cdot}040, 4/300)$$

$$\text{Uninformative posterior} = N(0{\cdot}040, 4/300)$$

$$\text{Sceptical posterior} = N(0{\cdot}028, 4/429)$$

$$\text{Enthusiastic posterior} = N(0{\cdot}115, 4/429).$$

Probabilities of improvement being greater than:

Target improvement	Log hazard	Uninformative	Sceptical	Enthusiastic
0·000	(0·000)	0·636	0·614	0·884
0·050	(0·149)	0·172	0·105	0·362
0·100	(0·290)	0·015	0·003	0·035

Interpretation: Adopting the stance of an enthusiast, we find that there is only a small chance (36·2 per cent) that there could be a difference of 5 per cent or greater. Someone with a more open mind might prefer to use the 'uninformative' prior, which suggests that the chance is even smaller, at 17 per cent. A sceptic, of course, would feel that a probability of 10 per cent tends to confirm his prior beliefs. The probability of a 10 per cent improvement is especially unlikely.

Recommendations: Even an enthusiast would have to agree that the chance of a treatment advantage of 10 per cent is exceedingly slim, and that there is but a small chance of a 5 per cent improvement. Probably it would be a waste of resources if more patients were entered into the trial. The trial should be discontinued.

However, there may be reasons beyond the calculations presented here for the trial to continue. For example, if another related trial has recently reported a large effect in favour of chemotherapy then it may be decided to continue the current trial so as to refute (or confirm) this other result. Although in principle this could be modelled by introducing the data (likelihood) for this second trial as well, and amending the posterior distribution, in practice it might be felt that the decisions should be reached by discussion and the application of informed value judgement by the Data Monitoring Committee. As always, other considerations, whether wider statistical ones or non-statistical, might lead to a decision other than that suggested by a formal statistical stopping rule.

18. CONCLUSION

Bayesian monitoring, as illustrated here, is very simple to implement. It helps to put into perspective one major inherent problem in the early termination of trials, namely the risk that the results will be regarded by sceptical clinicians as inconclusive. Thus it should help to ensure that trials only stop early when the results to date are sufficiently conclusive. Furthermore, our experience has been that in this context clinicians find Bayesian concepts intuitively appealing; the idea of collecting sufficient data to convince not only enthusiasts but both those with open minds, and sceptics too, accords with most clinicians' experience of research and the introduction of new drugs.

The methods described in this paper were initially tested retrospectively on the MRC randomized trials of misonidazole for head and neck cancer,[1] neutron therapy for pelvic cancer,[1] and chemotherapy for osteosarcoma.[20] Following the successful application of the methodology, the Bayesian approach for monitoring was adopted for prospective use on the two MRC trials of continuous hyperfractionated accelerated radiotherapy (CHART) for head and neck cancer and bronchus cancer,[1,2] and the on-going MRC oesophageal cancer trial (OE02) that provided basis of the hypothetical worked examples. Our experience to date leads us to recommend these procedures for more general application in monitoring of randomized controlled trials.

It is, perhaps, also worth finishing on a cautionary note. Bayesian analyses, like any other statistical calculations relating to stopping rules, provide information which should be considered within the more general context of the impact of early termination of a trial. There may be many reasons for deciding to ignore the calculations of posterior probabilities, and to continue recruiting patients to the trial. For example, there many be considerations concerning other endpoints such as toxicity, quality of life, and cost. Also, if a trial is stopped prematurely, some readers may regard the results as dubious and less convincing whatever approach is used; in effect, they may have an additional scepticism which they apply to all trials which are stopped early. Perhaps this trial may contribute to meta-analyses or overviews. If a trial is terminated permaturely, it may become difficult to later launch a confirmatory study. Thus one should never blindly and naively use Bayesian or any other monitoring in isolation and without full consideration of the implications.

Nevertheless, we believe that the Bayesian approach presented here makes explicit many of the issues involved in the monitoring of trials, and because of this it deserves to be more widely used.

ACKNOWLEDGEMENTS

This paper was motivated by the MRC Data Monitoring Committee for the oesophageal trial (OE02), and we would like to thank the clinical members Professor W. Duncan and Dr. J. Dark, and the clinical coordinator Dr. D. J. Girling.

REFERENCES

1. Spiegelhalter, D. J., Freedman, L. S. and Parmar, M. K. B. 'Bayesian approaches to randomised trials', *Journal of the Royal Statistical Society, Series A*, **157**, 357–416 (1994).
2. Parmar, M. K. B., Spiegelhalter, D. J. and Freedman, L. S. 'The CHART trials: Bayesian design and monitoring in practice', *Statistics in Medicine*, **13**, 1297–1312 (1994).
3. Abrams, K., Ashby, D. and Errington, D. 'Simple Bayesian analysis in clinical trials: a tutorial', *Controlled Clinical Trials*, **15**, 349–359 (1994).
4. Freedman, L. S. and Spiegelhalter, D. J. 'Application of Bayesian statistics to decision making during a clinical trial', *Statistics in Medicine*, **11**, 23–25 (1992).
5. Freedman, L. S. and Spiegelhalter, D. J. 'Comparison of Bayesian with group sequential methods for monitoring clinical trials', *Controlled Clinical Trials*, **10**, 357–367 (1989).
6. Freedman, L. S. and Spiegelhalter, D. J. 'The assessment of subjective opinion and its use in relation to stopping rules for clinical trials', *Statistician*, **32**, 153–160 (1983).
7. Pocock, S. J. *Clinical Trials: A Practical Approach*, Wiley, Chichester, 1983.
8. Pocock, S. J. 'Interim analyses for randomised clinical trials: the group sequential approach', *Biometrics*, **38**, 153–162 (1982).
9. Fayers, P. M., Cook, P. A., Machin, D., Donaldson, N., Whitehead, J., Ritchie, R., Oliver, R. T. D. and Yuen, P. 'On the development of the Medical Research Council trial of alpha-interferon in metastatic renal carcinoma', *Statistics in Medicine*, **13**, 2249–2260 (1994).
10. ICRF Clinical Trial Service Unit. *ISIS 3 Protocol*, CTSU, Radcliffe Infirmary, Oxford UK, 1989.
11. Parmar, M. K. B. and Machin, D. 'Monitoring clinical trials – experience of and proposals under consideration by the Cancer Therapy Committee of the British Medical Research Council', *Statistics in Medicine*, **12**, 495–504 (1993).
12. O'Brien, P. C. and Fleming, T. R. 'A multiple testing procedure for clinical trials', *Biometrics*, **35**, 549–556 (1979).
13. Berry, D. A. 'A case for Bayesianism in clinical trials', *Statistics in Medicine*, **12**, 1377–1393 (1993).
14. Fayers, P. M. and Machin, D. 'Sample size: how many patients are necessary?', *British Journal of Cancer*, **72**, 1–9 (1995).
15. Machin, D. and Campbell, M. J. *Statistical Tables for the Design of Clinical Trials*, Blackwell Scientific Publications, Oxford, 1987.

16. Machin, D., Campbell, M. J., Fayers, P. M. and Pinol, A. *Sample Size Tables for Clinical Studies*, Blackwell Science, Oxford, 1997.
17. Ashby, D. 'Preface: Papers from the conference on Methodological and Ethical Issues in Clinical Trials', *Statistics in Medicine*, **12**, 1373–1374 (1993).
18. Grossman, J., Parmar, M. K. B., Spiegelhalter, D. J. and Freedman, L. S. 'A unified method for monitoring and analysing controlled trials', *Statistics in Medicine*, **13**, 1815–1826 (1994).
19. Tsiatis, A. A. 'The assymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time', *Biometrika*, **68**, 311–315 (1981).
20. Spiegelhalter, D. J., Freedman, L. S. and Parmar, M. K. B. 'Applying Bayesian thinking in drug development and clinical trials', *Statistics in Medicine*, **12**, 1501–1511 (1993).

3.3 Analysis

TUTORIAL IN BIOSTATISTICS
LONGITUDINAL DATA ANALYSIS (REPEATED MEASURES) IN CLINICAL TRIALS

PAUL S. ALBERT*

Biometric Research Branch, National Cancer Institute, CTEP, DCTDC Executive Plaza North, 6130 Executive Blvd, MSC 7434 Bethesda, MD 20892-7434, U.S.A.

SUMMARY

Longitudinal data is often collected in clinical trials to examine the effect of treatment on the disease process over time. This paper reviews and summarizes much of the methodological research on longitudinal data analysis from the perspective of clinical trials. We discuss methodology for analysing Gaussian and discrete longitudinal data and show how these methods can be applied to clinical trials data. We illustrate these methods with five examples of clinical trials with longitudinal outcomes. We also discuss issues of particular concern in clinical trials including sequential monitoring and adjustments for missing data. A review of current software for analysing longitudinal data is also provided. Published in 1999 by John Wiley & Sons, Ltd. This article is a US Government work and is the public domain in the United States.

1. INTRODUCTION

The defining feature of a longitudinal study is that individuals are measured repeatedly across time. This is in contrast to cross-sectional studies where observations are taken at only one fixed point in time. In this tutorial, I will make the distinction between repeated measures and survival analysis where individuals are followed and their time to event or censoring is analysed. Our focus will be on methods for analysing longitudinal studies where the outcome variable is observed repeatedly in time. Longitudinal data are collected in clinical trials for various reasons. First, these data can be collected to obtain a more precise estimate of the outcome and hence the treatment effect; for example, in a blood pressure trial repeated blood pressure measurements close in time may be collected. Second, longitudinal data may be collected for monitoring purposes, although the primary focus is on the evaluation of the treatment effect at a particular time post-randomization. Third, longitudinal data may be collected to evaluate the effect of treatment over time. This tutorial will review methodology for analysing longitudinal data when interest focuses on this third reason. There have been many books and papers reviewing the literature on methodological developments in longitudinal data analyses.[1-7] There has been no thorough review from the perspective of designing and analysing data in clinical trials. The aim of this paper is to synthesize much of the research in longitudinal data analysis from the viewpoint of someone conducting and analysing the data from clinical trials.

* Correspondence to: Paul S. Albert, Biometric Research Branch, National Cancer Institute, CTEP, DCTDC Executive Plaza North, 6130 Executive Blvd, MSC 7434 Bethesda, MD 20892-7434, U.S.A.

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
 ISBN: 0-470-02365-1

I will discuss five clinical trials where the primary outcome is observed repeatedly in a longitudinal setting. The scientific focus in all these studies is on evaluating the effect of treatment over time. These clinical trials taken from a wide range of medical areas will serve as motivation for the methodological issues relevant to analysing longitudinal data from clinical trials. These examples are:

1. The Intermittent Positive Pressure Breathing (IPPB) Trial.[8] This was a controlled clinical trial evaluating the effect of an intervention (IPPB) as compared with the standard compressor nebulizer therapy on pulmonary function over time in 985 patients with obstructive pulmonary disease. The primary response was forced expiratory volume in one second (FEV_1) measured at baseline and at 3-month intervals over a 3 year follow-up period.
2. The Dietary Intervention in Children (DISC)[9] study. The randomized intervention study was designed to evaluate the efficacy and safety of a lipid-lowering diet on 663 children 8 to 10 years old with moderately elevated low-density-lipoprotein cholestrol (LDL-C) levels. The primary outcome was LDL-C measured at baseline and at yearly intervals over a 3 year follow-up period. Secondary outcomes were physical and cognitive growth and sexual maturation which were all collected longitudinally.
3. Alpha-interferon trial and chronic progressive multiple sclerosis (MS).[10] This was a clinical trial on 100 subjects which examined the effect of alpha-interferon on the course of multiple sclerosis. The primary outcome was the expanded disability status scale (EDSS), an ordinal scale of disease severity ranging from 0 to 10, measured at baseline, 1, 6, 12, 18, 24 and 36 months post-randomization. Secondary outcomes were magnetic resonance imaging (MRI) data collected longitudinally. In addition, longitudinal data on the occurrence of clinical exacerbations (defined as a new or worsened clinical disability persisting for at least 24 hours) was also collected.
4. An epilepsy clinical trial where 59 patients were randomized to either Progabide or placebo and followed longitudinally.[11,12] Data consisted of a baseline seizure count recorded over an 8-week period and four consecutive seizure counts observed over 2-week intervals post-randomization.
5. A randomized clinical trial for treating drug addiction where two treatments (Buprenorphine and Methadone) were compared for their ability to reduce opiate use among a group of 162 addicts.[13] The outcome of this trial is a vector of repeated binary responses of whether an individual failed a urine test at each of 3 visits per week (on Monday, Wednesday and Friday) over a 17-week period.

Methods discussed in this paper will be illustrated with the above examples. The remainder of the paper will be outlined in the following sections. Section 2 will discuss longitudinal methods for Gaussian data. We will discuss simple univariate, multivariate and random-effect approaches for analysing these data. Section 3 will discuss longitudinal methods for discrete data. I will review models for marginal inferences, random-effects models and transitional models. Section 4 discusses longitudinal methods for the analysis of recurrent events. Work on the problem of missing data in longitudinal studies will be discussed in Section 5. Work on sequential monitoring in longitudinal studies will be discussed in Section 6. Available software for longitudinal studies will be presented in Section 7. In the Appendix, I present S-plus code for analysing: (i) the IPPB trial data with Gaussian random effects models; (ii) the epilepsy clinical trial data with generalized estimating equations. A discussion which focuses on some additional topics and future directions

will be discussed in Section 8. Throughout this tutorial, I will draw upon our motivating examples when discussing each of the topics

2. LONGITUDINAL METHODS FOR GAUSSIAN DATA

Longitudinal Gaussian outcomes are common in clinical trials. The primary interest in the IPPB trial, for example, is the change in lung volume function (measured by FEV_1) over time. The study was designed to examine the average change in slopes across the treatment arms. Various approaches have been proposed for modelling longitudinal Gaussian data. Approaches can be divided into four major categories: simple univariate methods; multivariate methods including traditional growth curve modelling and generalized estimating equations (GEE), and random-effects models.

2.1. Univariate Methods

One simple approach is to summarize each person's longitudinal observations (for example mean, median, maximum value, last value or estimated slope) and compare these univariate measures across treatment groups.[14] These comparisons can be made using univariate statistical techniques (for example, t-test or Wilcoxon test) since there is only one summary measure per person and observations on different people are independent. In the IPPB trial, for example, the progression of lung function over time was compared by testing differences in the means of individual estimated slopes between treatment groups.[8] A univariate comparison of slopes can be an attractive alternative to more complex modelling methodology. In highly imbalanced data, however, such as the IPPB study where the number of repeated observations ranged from 2 to 12, an unweighted univariate comparison can be inefficient.

2.2. Multivariate Methods

In this category we include traditional multivariate and growth curve analysis techniques and generalized estimating equations. Traditional multivariate methods[15] such as the Hoteling's T^2, the analogue of the t-test for testing whether mean vectors are different in two samples, and profile analysis, a method for testing for parallelism and differences between two mean vectors in multivariate data, can be used to analyse continuous longitudinal data when observations are taken at the same time points on all subjects. Traditional growth curve modelling has focused on flexible comparisons of multivariate mean vectors[16–19] with unspecified correction structures. These methods are very useful in analysing clinical trials data when the primary outcome is Gaussian, is observed at regularly spaced intervals, and is rarely missing.

Others have proposed methods for analysing repeated longitudinal data using linear models with correlated errors. Denote Y_{ij} and $\boldsymbol{X}_{ij}$ as the Gaussian outcome and a p by 1 vector of covariates for the jth observation taken on the ith subject, respectively. A model with time-series error structure can be formulated as $Y_{ij} = \boldsymbol{X}_{ij}'\boldsymbol{\beta} + \varepsilon_{ij}$, where the ε_{ij} have a time-series structure. Here, $\boldsymbol{\beta}$ measures the linear effect of a set of covariates on the longitudinal response over time. For clinical trials, these covariates typically include treatment group, time since randomization, and treatment by time interactions. Various methods have been proposed[20–22] that incorporate a time-series error structure. An autoregressive moving average (ARMA) model is a common

choice for the error structure. For first-order dependence this general class of models reduces to

$$\varepsilon_t = \theta\varepsilon_{t-1} + \gamma a_{t-1} + a_t$$

where $a_t \sim$ i.i.d. $N(0, \sigma_\varepsilon^2)$, and θ and γ are the first-order autoregressive and moving-average effects, respectively. The linear model with an autoregressive error structure fits into the generalized estimating equations (GEE) model framework. The GEE[23,24] approach is a framework for modelling continuous and discrete longitudinal data which extends the generalized linear model[25] from the independent observation case to the repeated longitudinal data setting. The methodology is most useful for analysing discrete longitudinal data, but for continuous outcomes reduces to more standard linear models with correlated error terms. We discuss GEE in more detail in Section 3.

A common feature of all these multivariate models is that we explicitly model the correlation structure (unspecified or time series structure). Although there are some exceptions, most of these methods do not handle large amounts of missing and irregularly spaced observations. In the next section, we propose classes of models which are particularly ammendable to these complications.

2.3. Random-effects Models

Random-effects models are an alternative to multivariate methods for the analysis of repeated longitudinal data. They provide the basic modelling framework for much of the clinical trials related methodology that we will discuss later in this tutorial. These models are particularly useful for analysing longitudinal clinical trials in which there is a sizable number of missing observations (either due to missed visits, loss to follow-up, or death). In the IPPB trial, for example, although follow-up visits were scheduled at 3-month intervals, many follow-up visits were taken at intermediate time points due to scheduling problems or were missing due to death or drop-out; only 77 out of 985 patients (8 per cent) had complete equally spaced follow-up measurements. Our discussion will begin with linear random-effects models, followed by more recent non-linear approaches.

2.3.1. Linear Models

The Laird and Ware[26] model has become seminal work in this area. Cnaan *et al.*[27] provide a detailed review of these methods with an application to a schizophrenia clinical trial. The model is typically formulated in two stages. Conditional on each subject's random effects, the model for the first stage is

$$Y_{ij} = \boldsymbol{X}_{ij}'\boldsymbol{\beta} + \boldsymbol{Z}_{ij}'\gamma_i + \varepsilon_{ij}$$

where $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_{ij}$ are p and q element vectors of fixed- and random-effects covariates, respectively, and ε_{ij} are independent Gaussian random variables with mean 0 and variance σ_ε^2. In the second stage, the random effects, γ_i, are assumed to have a multivariate normal distribution with mean $\boldsymbol{0}$ and variance-covariance matrix $\boldsymbol{D}$. The parameter vector $\boldsymbol{\beta}$ and γ_i measure the effects of the fixed- and random-effect covariates, respectively, on the mean response. In a clinical trial, fixed-effect covariates may include treatment group, time since randomization, treatment by time interactions, and baseline characteristics, while typically only the intercept and time-effect are included as random-effect covariates. Group changes can be estimated with fixed-effect parameters, while an individual's departure from group averages can be assessed by estimating random effects. Various authors have discussed estimation for these models. Laird and Ware[26]

among others[28] have proposed iterative estimation procedures, while Vonesh and Carter[29] have proposed a non-iterative estimation procedure. There has been additional work in model diagnostics[30–32] and assessing whether the random-effects distribution is normal.[33] Although Butler and Louis[34] demonstrate that the random-effects distribution has little effect on fixed-effects estimation, recent work by Verbeke and Lesaffre[35] demonstrated the sensitivity of random-effects estimation to the distributional assumption on the random effects. These results emphasize the importance of model diagnostics for the random-effects distribution. One limitation of the general random-effects formulation is the assumption in the first stage that the error-structure conditional on the random effects is independent. This may be a poor assumption in certain applications. Chi and Reinsel[36] proposed a class of models with both random effects across individuals and autoregression in the within-individual errors. In addition, they develop a test for detecting autocorrelation in the within-individual errors.

2.3.2. Linear Model: Worked Example

The IPPB study provides good motivation for this methodology. A major interest is understanding the effect of treatment on the change in lung function (FEV_1) over time. The trial was designed to test for a change in average slope between the experimental and standard treatment groups at the end of the study. Random-effects models provide an attractive framework for analysing the data from this trial.

A basic model is

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 T_i + \beta_3 T_i t_{ij} + \gamma_{i0} + \gamma_{i1} t_{ij} + \varepsilon_{ij} \tag{1}$$

where $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ and $\gamma_i = (\gamma_{i0}, \gamma_{i1})' \sim$ i.i.d. $N(0, \boldsymbol{D})$, T_i is equal to 1 when the ith subject randomized to IPPB treatment and equal to 2 when randomized to standard compressor nebulizer therapy, and t_{ij} is the time from randomization for the jth follow-up time on the ith subject. The coefficient β_1 measures change over time, β_2 measures the treatment differences at baseline, and β_3 the difference of treatments over time. The two random effects γ_{i0} and γ_{i1} reflect individual departure in baseline FEV_1 measurements and slope, respectively. In addition, the diagonal elements in the random-effects variance, $\mathbf{D}$, summarize the between-subject variation in baseline and slope measurements.

We transformed the outcome variable to the log-scale since the outcome was close to being normally distributed on that scale. We fit this model using the linear mixed models routine in S-plus; computer code for this model is supplied in the Appendix. The parameter estimates were: $\hat{\beta}_0 = 0{\cdot}00644$ (SE = 0·378); $\hat{\beta}_1 = -0{\cdot}00359$ (SE = 0·000773); $\hat{\beta}_2 = -0{\cdot}0203$ (SE = 0·0241), and $\hat{\beta}_3 = 0{\cdot}00014$ (SE = 0·00049). The test of whether β_3 is zero provides a test of treatment effect; the z-value was 0·28, suggesting that there is no effect of treatment on change in FEV_1 over time. Because of the randomization, we would expect no difference in baseline FEV_1 by treatment group; the z-value corresponding to a test of β_2 equal to zero was $-0{\cdot}84$. The standard errors of the intercept and slope random effects (γ_{i0} and γ_{i1}, respectively), computed as the square root of the diagonal elements of $\mathbf{D}$ were 0·362 and 0·00508, respectively. The magnitude of these values suggests that there is substantially more between-subject variation in baseline FEV_1 measurements than in their change over follow-up time.

In order to show the advantages of this approach over a simple comparison of individually computed least-squares estimates of slope, I compared the differences in the slope estimates by

treatment group. The difference in average slope by the simple method was computed as 0·00006 (SE = 0·0001). We compare this value to our estimate of $\hat{\beta}_3$. Although both approaches result in insignificant effects, the resulting standard error in the simple two-stage approach was twice as high compared with the result obtained with the random-effects approach.

Adjustments for baseline covariates can easily be made by including them as independent variables in the model. Subgroup analyses can be performed by including a higher-order interaction between time, treatment group and the subgroup indicator as a fixed effect in the model. For example, in the IPPB trial we can test for a differential treatment effect by gender with the model

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 T_i + \beta_3 \text{Sex}_i + \beta_4 T_i t_{ij} + \beta_5 T_i \text{Sex}_i$$

$$+ \beta_6 \text{Sex}_i t_{ij} + \beta_7 T_i t_{ij} \text{Sex}_i + \gamma_{i0} + \gamma_{i1} t_{ij} + \varepsilon_{ij}$$

where a test of β_7 equal to 0 provides a test of a differential effect of treatment by gender. The z-value for this test was 0·119, suggesting that the treatment effect on the rate of change in FEV_1 was not different by gender.

The linear assumption was very reasonable for comparing the changing FEV_1 over time in this study; in each treatment group, non-parametric smoothed curves were closely approximated by linearity. In other studies the linear assumption may not be reasonable. In the DISC study, for example, each subject's height and weight are measured at yearly intervals. One option is to fit piecewise linear random-effects models (that is a model with a fixed change-point in the slope). Another possibility is to fit non-linear mixed models. We discuss these models in the next section.

2.3.3. Non-linear Models

One of the secondary interests in DISC was to evaluate the effect of a lipid-lowering diet on growth and development in children. Linear or piecewise linear models may inadequately describe growth variables like height. Non-linear mixed models have been proposed [37-39] that would be of particular interest in this application. Morrell *et al.*[40] describe an application of this methodology to modelling longitudinal cancer marker data. The non-linear mixed model, as with its linear analogue is formulated in two stages. In the first stage, a class of non-linear curves with fixed and subject-specific random effects describe each individual's curve. In the second stage, a distribution on the random effects ties together how each individual's curve differs from an overall curve. Conditional on each subject's random effects, Lindstrom and Bates[37] propose the following first-stage model:

$$Y_{ij} = f(\boldsymbol{\phi}_i, \boldsymbol{x}_{ij}) + \varepsilon_{it}$$

$\boldsymbol{\phi}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\gamma_i$, $\boldsymbol{\beta}$ and γ_i are vectors of fixed and random effects, $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ are design matrices for the fixed and random effects, respectively, f is a non-linear function which is a function of $\boldsymbol{\phi}_i$ and a vector of covariates $\boldsymbol{x}_{ij}$, and $\varepsilon_{ij} \sim \text{N}(0, \sigma_\varepsilon^2)$. As in the linear mixed model, γ_i are assumed to have a multivariate normal distribution with mean 0 and variance-covariance matrix $\boldsymbol{D}$. An example of a model which may adequately describe growth for DISC subjects is the logistic growth model

$$Y_{ij} = f(\boldsymbol{\phi}_i, \boldsymbol{x}_{ij}) + \varepsilon_{ij} = \frac{\phi_{i1}}{1 + \phi_{i2}\exp(\phi_{i3} t_{ij})} + \varepsilon_{ij}$$

where t_{ij} is the time from randomization at the jth visit for the ith subject. In this formulation, an individual's asymptote is ϕ_{i1}, while the progression to that asymptote is governed by ϕ_{i3}. We can examine the effect of treatment (denoted by the indicator variable T_i) on these two important characteristics of these individual curves by the following decomposition of the fixed and random effects, $\phi_{i1} = \beta_0 + \beta_1 T_i + \gamma_{i1}$, $\phi_{i2} = \beta_2 + \gamma_{i2}$, and $\phi_{i3} = \beta_3 + \beta_4 T_i + \gamma_{i3}$. In Lindstrom–Bates's notation, $\boldsymbol{\phi}_i = (\phi_{i1}, \phi_{i2}, \phi_{i3})'$, $x_{ij} = t_{ij}$,

$$A_i = \begin{pmatrix} 1 & T_i & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & T_i \end{pmatrix} \qquad B_i = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)'$ and $\gamma_i = (\gamma_{i1}, \gamma_{i2}, \gamma_{i3})'$ for this example.

We can test for the effect of treatment on the asymptote and the rate of reaching that asymptote by evaluating the significance of β_1 and β_4.

Lindstrom and Bates propose an iterative EM-type algorithm for parameter estimation. Vonesh and Carter[38] propose a non-iterative estimation procedure which does not assume Gaussian random effects for a slightly different model formulation. Vonesh *et al.*[41] propose goodness-of-fit tests for the adequacy of the random effects in describing the correlation structure.

3. LONGITUDINAL METHODS FOR DISCRETE DATA

Many longitudinal outcomes in clinical trials are binary, counts or categorical. Each of our examples have a non-Gaussian longitudinal response either as a primary or secondary outcome. Many models have been proposed for analysing discrete longitudinal data. They can generally be divided into one of three types of approaches: marginal, random effects, and transitional models. We will discuss each in turn.

3.1. Marginal Models

Marginal models focus on estimating the effect of a set of covariates on the marginal expectation of the response. In the epilepsy trial, for example, we are interested in comparing average seizure rates across treatment over time while adjusting for a set of baseline covariates. Liang and Zeger's[23,24] seminal work provided the general framework for these models. Their initial formulation extends the generalized linear model[25] to adjust for correlation on observations on the same subject. Denote μ_{ij} as the mean of the jth response on the ith subject (that is, $\mu_{ij} = E[Y_{ij}]$). In the terminology of generalized linear models, the mean is related to a set of covariates through a link function. h, where $h(\mu_{ij}) = h(E(Y_{ij})) = \boldsymbol{X}_{ij}'\boldsymbol{\beta}$. In addition, the functional relationship between the variance and mean is specified as $\text{var}(Y_{ij}) = \phi g(\mu_{ij})$ and the correlations on observations taken on the same subject are characterized by the function $\text{corr}(Y_{ij}, Y_{ij'}) = \rho(\boldsymbol{\alpha})$. The approach is particularly attractive in that a general class of response variables can be modelled by choosing h and g in an appropriate way. For Gaussian longitudinal data like FEV_1 in the IPPB study, $h(\mu_{ij}) = \mu_{ij}$ and $g(\mu_{ij}) = 1$. For binary longitudinal data, as in the drug addiction trial, h and g are chosen as $h(\mu_{ij}) = \text{logit}(\mu_{ij})$ and $g(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$. Similarly, for repeated count data as in the epilepsy clinical trial, $h(\mu_{ij}) = \log(\mu_{ij})$ and $g(\mu_{ij}) = \mu_{ij}$. Inferences on $\boldsymbol{\beta}$ with the identity link, logistic link and log-link have the same interpretation as in multiple

regression, logistic regression and Poisson regression for Gaussian, binary and count data, respectively.

Liang and Zeger propose an estimating equations approach for parameter estimation which extended the concept of quasi-likelihood[25] to correlated observations. In particular, they propose solving the following set of estimating equations (called generalized estimating equations (GEE)):

$$\sum_{i=1}^{n} \boldsymbol{D}_i'(\boldsymbol{\beta})\boldsymbol{V}_i(\boldsymbol{\beta}, \alpha)^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = 0$$

for $\boldsymbol{\beta}$, where $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$, $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{in_i})'$, $\boldsymbol{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$, $\boldsymbol{V}_i = \text{var}(\boldsymbol{Y}_i) = \phi\boldsymbol{A}_i^{1/2}R(\boldsymbol{\alpha})\boldsymbol{A}_i^{1/2}$, $\boldsymbol{A}_i = \text{diag}(g(\mu_{i1}), g(\mu_{i2}), \ldots, g(\mu_{in_i}))$, n is the number of subjects, and where n_i is the number of observations on the ith subject. Liang and Zeger proposed various models for the correlation structure including an independent model, where

$$R(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and an exchangeable correlation structure, where

$$R(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{pmatrix}$$

and an unspecified correlation structure, where

$$R(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 \\ \alpha_1 & 1 & \alpha_3 \\ \alpha_2 & \alpha_3 & 1 \end{pmatrix}.$$

They propose a modified Fisher scoring algorithm for solving the estimating equations

$$\widehat{\boldsymbol{\beta}_{p+1}} = \widehat{\boldsymbol{\beta}_p} + \left(\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i(\boldsymbol{\alpha})^{-1}\boldsymbol{D}_i\right)^{-1} \sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i(\boldsymbol{\alpha})^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i), p = 1, 2, \ldots,$$

and thus estimating $\boldsymbol{\beta}$. Method-of-moment estimates for the correlation structure parameters $\boldsymbol{\alpha}$ and the scale parameter ϕ are proposed. Liang and Zeger demonstrate that $\boldsymbol{\beta}$ is consistent even when the correlation structure is misspecified. Model-based estimates of the variance of $\boldsymbol{\beta}$, given by

$$\text{var}_{\text{mod}}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i(\boldsymbol{\alpha})^{-1}\boldsymbol{D}_i\right)^{-1}$$

are inconsistent if the model for the correlation structure is misspecified. Liang and Zeger propose a robust estimator of variance which is consistent even when the model is misspecified. This estimator is

$$\text{var}_{\text{rob}}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i(\boldsymbol{\alpha})^{-1}\boldsymbol{D}_i\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)'(\boldsymbol{Y}_i - \boldsymbol{\mu})\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)\left(\sum_{i=1}^{n} \boldsymbol{D}_i'\boldsymbol{V}_i(\boldsymbol{\alpha})^{-1}\boldsymbol{D}_i\right)^{-1}.$$

Inference on $\boldsymbol{\beta}$ using the robust estimator of variance is asymptotically valid under any assumed correlation structure. An appropriate model for the correlation structure, however, results in more efficient estimation (particularly for time-dependent covariates such as a treatment-by-time interaction[42]). In addition, it has been noted[43] that the robust estimator may have poor finite sample properties in cases where there are more than a few correlated observations on each subject. Hypothesis testing for these models has been proposed by Rotnitzky and Jewell.[44]

Prentice[45] proposed the use of a second estimating equation for the correlation parameters $\boldsymbol{\alpha}$. In particular, for longitudinal binary data he proposed simultaneously solving an estimating equation for the mean structure parameters $\boldsymbol{\beta}$ as well as the following estimating equation for the correlation parameter $\boldsymbol{\alpha}$:

$$\sum_{i=1}^{n} \boldsymbol{E}_i' \boldsymbol{W}_i^{-1} (\boldsymbol{Z}_i - \boldsymbol{v}_i) = 0$$

where $\boldsymbol{Z}_i$ is a vector of the $n_i(n_i - 1)/2$ squared residuals corresponding to the unique off-diagonal elements of the variance-covariance matrix $\boldsymbol{V}_i$, $\boldsymbol{v}_i = E(\boldsymbol{Z}_i)$, $\boldsymbol{E}_i = \partial \boldsymbol{v}_i / \partial \boldsymbol{\alpha}$ and $\boldsymbol{w}_i = \mathrm{Var}(\boldsymbol{Z}_i)$ for the ith subject. This approach results in efficiency gains over the original formulation and allows for richer classes of models for the correlation structure.

Various authors[46,47] have proposed extensions of GEE to incorporate assumptions about higher-order moments. These methods have been called GEE-2 methods in contrast to the previous methods that have been referred to as GEE-1 methodology. In this approach the single joint estimating equation

$$\sum_{i=1}^{n} \left(\frac{\partial (\boldsymbol{\mu}_i, \boldsymbol{v}_i)'}{\partial (\boldsymbol{\beta}, \boldsymbol{\alpha})} \right) \mathrm{var}(\boldsymbol{Y}_i, \boldsymbol{Z}_i)^{-1} \begin{pmatrix} \boldsymbol{Y}_i - \boldsymbol{\mu}_i \\ \boldsymbol{Z}_i - \boldsymbol{v}_i \end{pmatrix} = 0$$

is solved for both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. The major advantage of GEE-2 over GEE-1 methodology is in efficiency gain. This efficiency gain is small for estimating $\boldsymbol{\beta}$ (generally 10 per cent or less), but can be quite large for estimating $\boldsymbol{\alpha}$.[47] Although GEE-2 methodology has been very useful in applications where inferences on the association structure are of primary interest,[48] they are less useful when inferences on the mean structure (as in most clinical trials) are of primary interest. Their major disadvantage is that unlike GEE-1, where inferences on $\boldsymbol{\beta}$ are valid even when the correlation structure is misspecified, inferences on $\boldsymbol{\beta}$ with GEE-2 require the correct model for the correlation structure.

Many clinical trials have a repeated categorical or ordinal response as a primary outcome. For example, the primary outcome for the MS alpha-interferon trials was an ordinal measure of disease severity measure at 7 time points over a 3-year follow-up. Generalized estimating equation approaches have been developed for categorical and ordinal longitudinal data. Miller *et al.*[49] demonstrated the connection between a GEE approach with a saturated correlation structure and a weighted least squares approach for modelling correlated categorical data. Lipsitz *et al.*[50] proposed modelling the correlation structure with a GEE approach, while Heagerty and Zeger[51] proposed a general class of models with improved efficiency.

Recently, there has been work in sample size estimation and in assessing model adequacy for GEE. Liu and Liang[52] propose formula derived expressions for sample size and power that are useful for designing clinical trials. Key references for assessing model fit include Preisser and Qaqish[53] who assessed the stability of mean structure parameter estimation, Albert and McShane[43] who assessed the correlation structure using variograms, Heagerty and Zeger[54] who

proposed a new diagnostic for assessing the correlation structure in longitudinal categorical data, and Barnhart and Williamson[55] who proposed a goodness-of-fit test for longitudinal repeated data. In general, it is difficult to assess model adequacy in GEE. Although likelihood-based methods are less robust than GEE, they are generally more efficient (only slightly for mean structure covariates) and provide a more well-defined means of assessing model adequacy. Various authors have proposed likelihood-based models for discrete data.[56-58]

Various authors have fit GEE models to discrete longitudinal clinical trial data. Diggle *et al.*[1] employ this method to demonstrate little effect of Progabide on seizure occurrence in the epilepsy clinical trial. S-plus code is presented for duplicating this analysis in the Appendix of this tutorial.

With the advent of GEE, the use of marginal models for discrete longitudinal data has become widespread. Some have argued[59] that these methods are most useful for observational studies, and less useful in clinical trials where assessing the effect of treatment on an average patient is of primary interest. Random-effects and transitional models have this interpretation. I will discuss them in the next two sections.

3.2. Random-effects Models

The basic idea in these models is that patient to patient variability is introduced by adding random effects as linear predictors in the regression relationship. Thus, in these models, heterogeneity and induced correlation can be thought of as arising from unobserved covariates. Zeger *et al.*[60] have called these 'subject-specified models' in contrast with 'population-averaged models' for marginal models. As with the linear and non-linear mixed models discussed in Section 2.3, these models can be viewed in two stages. In the first stage, the mean response for the ith person is $\mu_{ij} = h(Y_{ij}|\gamma_i) = \boldsymbol{X}_{ij}'\boldsymbol{\beta} + \boldsymbol{Z}_{ij}'\gamma_i$. In addition, the relationship between the mean and variance conditional on the random effects is specified as $\text{var}(Y_{ij}) = \phi g(\mu_{ij})$. In the second stage, a distribution for the random effects γ_i is postulated; often this is assumed normal with mean $\mathbf{0}$ and variance $\boldsymbol{D}$. This formulation encompasses a wide range of random-effects models. The linear mixed model discussed in Section 2 result when $h(\mu_{ij}) = \mu_{ij}$ and $g(\mu_{ij}) = 1$, a random-effects model for logistic regression results when the response is binary and $h(\mu_{ij}) = \text{logit}(\mu_{ij})$ and $g(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$, and a random-effects model for count data results when $h(\mu_{ij}) = \log(\mu_{ij})$ and $g(\mu_{ij}) = \mu_{ij}$.

Regression parameters can be interpreted as the effect of changing an explanatory factor on an individual's average response. Regression coefficients are generally different (with the exception of the identity link function used with Gaussian responses) from marginal model coefficients.[60,61] Tests of significance for individual coefficients are generally similar between the two approaches. However, tests for time-dependent covariates (for example, a covariate such as time since randomization in the epilepsy clinical trial) are usually more powerful with random-effects models than with marginal models.[62] Diggle *et al.*[1] analyse the Progabide epileptic seizure clinical trial data with both marginal and random-effects approaches and show stronger treatment by time interactions with the random-effects model as compared with the marginal model.

Various estimation procedures have been proposed. Stiratelli *et al.*[63] discussed an EM algorithm approach for binary response data with Gaussian random effects, and Longford[64] discussed an approach based on direct maximization of the likelihood. In addition, Zeger *et al.*[60] and Waclawiw and Liang[65] proposed methodology for estimating fixed and random effects in a GEE setting. Gibbs sampling[66] as well as pseudo-likelihood[67] approaches have also been proposed for

Gaussian random effects. In addition, methods that estimate the random-effects distribution non-parametrically[68] have been proposed.

3.3. Transitional Models

Transitional models examine the effect of covariates on the transition patterns across a binary and categorical response over time. In the drug addiction trial the effect of treatment on the beginning and continuation of opiate use is of primary interest. For the multiple sclerosis example, the effect of treatment on an individual's disease course is of primary interest. Here we may wish to examine the effect of treatment on the propensity to transition to EDSS levels with greater severity. Various authors have proposed autoregressive-type regression models for modelling transitional patterns in binary, categorical or ordinal longitudinal data. For binary data,[69-71] these models reduce to lagging previous observations in the mean structure

$$\text{logit}\, P(Y_{it} = 1 \mid Y_{it-1}, Y_{it-2}, \ldots, Y_{it-q}) = \boldsymbol{X}_{it}'\boldsymbol{\beta} + \sum_{l=1}^{q} \theta_l Y_{it-q}$$

where $\boldsymbol{X}_{it}$ are subject-specific and time-dependent covariates, and q is the order of Markov dependence. The regression coefficients can be interpreted as the effect of covariates on the probability of a binary event adjusting for the past history of the process.

The opiate clinical trial provides a good example of this modelling approach. The model

$$\text{logit}\, P(Y_{ij} = 1 \mid Y_{ij-1}) = \beta_0 + \beta_1 T_i + \beta_2 \text{Mon}_{ij} + \beta_3 \text{Wed}_{ij} + \beta_4 Y_{ij-1}$$

where T_i, is an indicator of whether the ith subject is on Buprenorphine (as opposed to the standard Methadone treatment) and Mon_{ij}, Wed_{ij} are indicators which incorporate a day of the week effect. The effect of treatment on the transition pattern can be assessed by examining β_1. For the opiate clinical trial, the estimate of β_1 was highly significant, suggesting that addicts on Buprenorphine have a lower propensity to enter and stay in a period of opiate-use than addicts on Methadone. A complete analysis of this data is presented in Albert.[72]

Extensions of these models to repeated categorical[73] and ordinal[1] data have also been considered. These models are in discrete time where observations are at regularly spaced intervals. The design and analysis of continuous-time transitional processes from data observed in discrete time has been considered.[74-76] These approaches allow for transitional inferences from highly irregularly spaced observations by making modelling assumptions which relate instantaneous probabilities of state transitions to discrete-time transition probabilities. Generally, transitional models assume no heterogeneity in the transitional processes across patients. This may be a poor assumption in clinical trials where a treatment may naturally induce heterogeneity in the transition process (that is, the treatment works better for some patients than others). Recently, two approaches that allow for the transition probabilities in binary processes to follow a random-effects distribution have been proposed.[77,78]

4. LONGITUDINAL METHODS FOR RECURRENT EVENTS

Many clinical trials compare recurrent events across treatment arms. For example, clinical exacerbations are compared in the MS clinical trial and seizure occurrence is compared in the epilepsy clinical trial. There are two basic modelling approaches. One approach is to model the

repeated times between recurrent events. The second approach is to model the number of events over a fixed follow-up interval. We discuss these two approaches in turn.

Various authors have proposed Poisson regression models with random effects for modelling recurrence data where the time between recurrent events is observed.[79-81] These models are attractive in that they allow for baseline covariate adjustments and for subgroup analyses by examining higher-order interactions with these covariates and treatment group indicators. Wei *et al.*[82] proposed methodology for multivariate failure time data which can also be applied in this setting. These has also been work in hypothesis testing; Cook *et al.*[83] proposed non-parametric tests for comparing the recurrence processes between two or more groups without modelling assumptions. These tests compare non-parametric estimates of cumulative mean functions across treatment groups. Cook[84] discussed design of clinical trials with recurrent events. He derived expected sample sizes and study duration to achieve a specified power under Poisson and over-dispersed Poisson models.

In some clinical trials it is not feasible to obtain reliable information on the exact times of recurrent events. Collecting the number of events over a fixed follow-up may be an attractive alternative. Poisson regression[25] may be used to analyse such frequency data. Recurrent events in clinical trials often have additional information associated with them. In the MS clinical trial clinical exacerbations have an associated measure of disease severity, while in the epilepsy study seizures are rated in terms of their type and severity. There has been recent work proposing methodology for combining recurrence and severity data. Barnhart and Sampson[85] proposed a likelihood-based approach which can be applied to simultaneously model recurrent events and associated continuous measures of event-specific severity. Albert *et al.*[86] generalized their methodology in jointly modelling the number of events and a vector of correlated discrete severity measures using a GEE approach. Let n_i be the number of events and $\boldsymbol{Y}_i | n_i$ as an n_i-dimensional vector of continuous or discrete severity measures, each element of which corresponds to one of the n_i events. This approach functionally links the event rate with the marginal mean of the severity measures through a shared parameter. In particular, the mean structure is modelled as

$$\log E(n_i) = \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{W}_i'\gamma_1 \quad h(E(Y_{ij} | n_i)) = \theta \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{U}_{ij}\gamma_2$$

where $\boldsymbol{X}_i$ is a p-dimensional vector of similar acting covariates, $\boldsymbol{\beta}$ is a vector of regression coefficients for these covariates, and θ is a scalar parameter that can be thought of as scaling the effect of these covariates on the count mean to the h link scale. In addition, $\boldsymbol{W}_i$ and $\boldsymbol{U}_{ij}$ are q- and r-dimensional vectors of additional covariates, and γ_1 and γ_2 are associated vectors of regression coefficients. An application to epilepsy clinical trial data is presented in Albert *et al.*[86]

5. ANALYSING MISSING DATA IN LONGITUDINAL CLINICAL TRIALS

Missing data is a major problem in longitudinal clinical trials. In the drug addiction trial more than half the subjects have missing observations by the fourth week of follow-up. The missing data in this study is due to both drop-out as well as intermittently missing data. In addition, the proportion of missed observations was different by treatment arm; patients randomized to the Methadone arm had higher proportions of missed observations than those randomized to Buprenorphine. This clearly has the potential to confound treatment effect assessment. The IPPB trial also had substantial amounts of missing data. (Approximately 23% of the patients died before the end of the study and 13.5% dropped out because they moved away or refused to return for follow-up visits.)

Patients may be missing for many reason including inability to comply with the treatment, a poor outcome, a good outcome, or the missing data mechanism may be totally random. All but the last reason may be a problem in most standard statistical methods for analysing longitudinal data. The effect of missing data on standard analyses depends on why the data are missing. Little and Rubin[87] classify missing data into three categories; this categorization will be useful to our discussion. First, data are missing completely at random (MCAR) if the missing data mechanism is independent of both the observed and actual missing value. In addition, data are missing at random (MAR) if the missing data mechanism depends on observed data but is independent of the actual missing value (for example, the missing data mechanism depends on previous responses through a logistic model). Last, data are missing non-randomly (also called non-ignorable missingness[87]) if the missing mechanism depends on the actual values of the missed observations. We discuss the impact of these types of missing data on various longitudinal data analysis methods.

The differences between the three different categories of missing data can be illustrated with the opiate clinical trial. MCAR would result if addicts missed their visits totally at random. A MAR missing data mechanism would result if the probability of missing a visit was directly related to prior observed responses. An example of a non-ignorable missing data mechanism would be if in addition to prior observed responses affecting the missingness, an addict would be more likely to miss a visit if they were taking opiates at the time than if they were not.

Generally speaking, any method that can handle unequally spaced and different numbers of observations on each subject can handle MCAR data. An attractive feature of the Laird–Ware linear-mixed model for Gaussian data (as opposed to standard multivariate data analysis techniques) is its facility for analysing such data. Other methods such as GEE can easily accommodate MCAR data.

Likelihood-based longitudinal methods easily accommodate MAR data. This follows by noting that the likelihood factors into the product of terms which depend on model parameters and terms which depend on missing data mechanism parameters. Thus methods such as the Laird–Ware model for Gaussian data and likelihood based methods for discrete and ordinal longitudinal data are not biased when data are MAR. Unfortunately, moment-based methods such as GEE are biased for MAR data.[88] Robins *et al.*[89] proposed an extension of GEE that allows for MAR data. They proposed a class of weighted estimating equations which result in consistent estimation of mean structure parameters with a correctly specified missing data mechanism. Their approach reduces to a weighted version of GEE where each element of the residual vector $(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$ is weighted by the inverse of the probability of having observed that response.

Both likelihood-based and moment-based methods are biased when there is non-ignorable missingness. Wu and Carrol[90] coin the term *informative* drop-out (or missingness) as a special type of non-ignorable drop-out (or missingness) which has received a lot of attention in the biostatistics literature. Wu *et al.*[90,91] and Mori *et al.*[92] proposed methodology for estimating the rate of change of a continuous repeated outcome when right censoring is informative. The need for this methodology is illustrated with the analysis of the primary outcome FEV_1 in the IPPB trial. Interest focuses on comparing the average slopes between treatment and placebo arms, adjusting for the differential censoring pattern due to death in the two groups. The basic idea in this work is that a shared random parameter is introduced between the repeated continuous outcome and the probability model for the censoring mechanism inducing a dependence. Estimation in the presence of these shared random effects adjusts for the informative missingness.

Follmann and Wu[93] generalized the previous model to account for informative missing data for generalized linear mixed models and applied their methodology to the addiction trial data. Their methodology is particularly attractive in that it allows for both informative drop-out as well as for informative intermittent missing data. Denote Y_{ij} as the jth response on the ith subject. We denote M_{ij} as indicators of whether the Y_{ij} response is missing. Follmann and Wu propose linking the longitudinal response data and the missing data mechanism through a shared random parameter

$$h(E(Y_{ij}|\gamma_i)) = \boldsymbol{X}'_{ij}\boldsymbol{\beta} + \boldsymbol{Z}'_{ij}\gamma_i$$

and

$$g(E(M_{ij}|\gamma_i)) = \boldsymbol{V}'_{ij}\boldsymbol{\eta} + \Delta'\gamma_i$$

where h and g are link functions which related the means of the responses and missing data indicators to sets of linear functions of covariates and random effects, $\boldsymbol{X}_{ij}$ and $\boldsymbol{V}_{ij}$ are covariates for the response and missing data mechanism, respectively, and $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are the associated parameter vectors. In addition, γ_i is a vector of random parameters common to both models, where the vector $\boldsymbol{Z}_{ij}$ governs how the random effects enter into the mean response and Δ is a parameter vector which characterizes the linkage between the response and missingness models. The shared random effects, γ_i, along with the parameter vector Δ characterize the informative missing mechanism. Follmann and Wu[93] propose an estimation procedure which makes fewer assumptions on the missing data mechanism by conditioning on the missing data pattern; their approach is termed an approximate conditional model. Their approach can be illustrated with opiate clinical trial data with the model

$$\text{logit}\, P(Y_{ij} = 1 \,|\, \gamma_i) = \beta_0 + \beta_1 T_i + \gamma_i$$

$$\textit{logit}\, P(M_{ij} = 1 \,|\, \gamma_i) = \eta_0 + \eta_1 T_i + \theta_1 T_i \gamma_i + \theta_2 (1 - T_i)\gamma_i$$

where a significant estimate of β_1 demonstrated that addicts on Buprenorphine had lower risk of opiate-use as compared with addicts on Methadone treatment, and a significant positive estimates of θ_1 and θ_2 adjusts for an informative missingness in which individuals with the most opiate-use have the highest propensity for missing data. Note that different values of θ_1 and θ_2 describe the interesting case in which the informative missingness mechanism is different in the two treatment arms.

Various authors have considered transitional models which adjust for non-ignorable missingness.[94,95] Albert[72] proposes a binary transitional model which incorporates both non-ignorable drop-out and intermittantly missingness. Using this approach, it is demonstrated that, even in the presence of non-ignorable missingness, the addicts on Bruprenorphine have a lower propensity to enter and stay in a period of opiate-use as compared with addicts on standard Methadone therapy.

As an alternative approach, Diggle and Kenward[96] proposed a model which combines a multivariate linear model for the underlying outcome with a logistic regression for modelling the drop-out process. Molenberghs *et al.*[97] extend this idea to repeated ordinal data. These approaches are marginal in contrast to the previously discussed random-effects and transitional models. In another work, Cook and Lawless[98] discussed methodology for analysing recurrent events with informative right censoring. They proposed a marginal approach which allows for easily interpretable inferences adjusting for this informative censoring.

Dawson and Lagakos[99] discussed the effect of missing data on comparing summary measures of longitudinal data across treatment arms. They proposed a test which stratifies over missing data pattern in order to preserve the proper type I error. Their approach makes few modelling assumptions and is particularly attractive in situations where the interest is on testing for an overall comparison across groups. Follmann *et al.*[100] discussed various imputation, model-based and combination tests on summary measures for testing treatment efficacy in clinical trials with repeated binary data and missed observations. They proposed tests of means and rank tests that are appropriate for the opiate addition trial. One approach they proposed is to construct a bivariate summary of each individual's response and missing data; for example, an average observed response and an average proportion missing. A global bivariate test of treatment effect simultaneously compares response and missing data mechanisms across treatment groups.

6. SEQUENTIAL MONITORING IN LONGITUDINAL CLINICAL TRIALS

Sequential monitoring in clinical trials in an important issue. Definitive clinical trials are usually set up to be monitored by an independent data safety and monitoring board (DSMB) which meets regularly to monitor the data for safety and efficacy. The board, which often meets at 3–6-month intervals over the length of the trial, must make a decision on whether the trial should be stopped at each meeting. The IPPB trial serves as an example. Longitudinal data (slopes) are being assessed at 3-month intervals, trends are examined, and the board is asked to make a judgement about efficacy or futility.

Various approaches have been proposed for monitoring clinical trials with univariate response data. The primary focus was to develop sequential boundaries for detecting treatment effect which adjust for the inherent multiplicity in examining the data repeatedly over the duration of the trial. Pocock[101] developed grouped sequential boundaries for the situation where the trial is monitored at equally spaced information times; the proposed boundaries were constant over the duration of the trial. O'Brien and Fleming[102] proposed boundaries which are monotonically decreasing for monitoring at equally spaced information times. Slud and Wei[103] proposed a technique for monitoring the trial at prespecified (before monitoring begins) unequally spaced information times. Lan and DeMetts[104] proposed methodology that allows for monitoring at unequal information times and does not require prespecifying the monitoring times. They proposed a spending function approach where a certain percentage of the type I error is 'spent at each look until it is all 'used up'.

There have been extensions of the group-sequential methods of Slud–Wei and Lan–DeMets to a few longitudinal models. Geary[105] developed sequential monitoring boundaries for repeated Gaussian data. This procedure assumes a basic four parameter model, that monitoring times are equally spaced, that patients enter the study all the same time and that there is no missing data. Lee and DeMets[106] proposed a method that does not require the previous assumptions and uses the Lan–DeMets spending function approach to compute exit probabilities. The approach assumes a Laird–Ware model for change in a continuous response where the estimate for linear change is being monitored. The method allows for the specification of a mechanism for staggered entry (that is, how patients enter into the trial over time). This is an important feature since it allows for distinguishing between the sequential and longitudinal components of the problem. For example, following a group of simultaneously enrolled subjects for 3 years provides a different amount of information than following a group of staggeredly entered patients an average of

3 years. Wu and Lan[107] further develop this approach by allowing for non-linear changes as well as informative censoring.

These methods could be applied to the IPPB trial if it were being conducted today. One would need to propose a model for the longitudinal decline in FEV_1 (for example, a linear mixed model), a detectable difference in slope between the two treatment arms, the variation in slopes across subjects, and the within-subject variation. The variance estimates can be obtained from prior studies. In addition, a staggered entry mechanism (this was uniform over a 3-year accrual period in the IPPB trial) and an α spending function (choices corresponding to Pocock and O'Brien–Flemming boundaries are given in Lan and DeMets[104]) needs to be specified. With this information, sequential boundaries for multiple analyses can be computed using a spending function approach.

Sequential monitoring plans have been constructed for GEE models. Wei *et al.*[108] have proposed sequential boundaries for GEE models that follow the prespecified 'looks' of Slud and Wei. Gange and DeMets[109] proposed a spending function approach for monitoring correlated data using GEE. Cook and Lawless discussed sequential monitoring for recurrent events.[110] They combined their non-parametric test for recurrent events[83] with Lan–DeMets spending function approach to develop a robust approach for monitoring recurrent events.

Stochastic curtailment is an alternative approach to the previous discussed methodology developed by Lan *et al.*[111] They proposed early stopping of a trial either for efficacy or futility based on conditional power. Specifically, at each monitoring time they proposed computing the power conditional on the available data and assuming a hypothesis for the future data. They proposed stopping the trial for efficacy when the conditional power is high and stopping the trial for futility when this power is low. Conditional power calculations for efficacy are often done assuming that future data is generated under the null hypothesis. The calculations for futility are performed assuming that future data are either generated similar to the data already observed or under another reasonable alternative hypothesis. McMahon *et al.*[112] discussed a simulation-based technique for stochastic curtailment in recurrent event studies. Lan and Zucker[113] discussed this idea applied to the basic single random-effects model for longitudinal Gaussian data. Halperin *et al.*[114] proposed stochastic curtailment methodology for the more general Laird–Ware random-effects model.

Halperin *et al.*'s methodology was motivated by the IPPB trial. In the IPPB trial, patients were randomized uniformly over a 3-year follow-up and each patient was followed for 3 years. Halperin *et al.* provide estimates of the conditional power of accepting the null hypothesis (that is, the probability that we will fail to reject the null hypothesis at the end of the study given the available data and that the minimal detectable differences is true) of 0·71, 0·83 and 0·90 at 3·5, 4·5 and 5·5 years after the beginning of the study. Based on this, stopping the IBBP trial at between 4·5 and 5·5 years might have been recommended.

Although sequential monitoring is useful in longitudinal studies, one has to be extremely cautious about stopping a trial just based on these statistical rules. Often, particularly in the early looks, linear effect estimates are based on short term data. It may be that long term effects are very different and this is most likely the reason for conducting a longitudinal study in the first place.

7. SOFTWARE FOR ANALYSING LONGITUDINAL DATA

Unfortunately, most published methodology for analysing longitudinal data has not been incorporated into commercial software. For most of the methods discussed in this paper one

needs to either develop new software or write to the authors for their research programs. Fortunately, there are some notable exceptions. Programs for fitting models to Gaussian data (that is, linear and non-linear mixed models and models with time-series error structure) and GEE for continuous and discrete data are available. SAS has a procedure called Proc Mixed[115] which fits linear models for Gaussian response data. It allows for time-series error structure (including unstructured and AR(1) error structure) and/or an arbitrary number of random effects (including the linear mixed model). The program provides maximum-likelihood estimation for the fixed effects, restricted maximum-likelihood for the variance components, and empirical Bayes estimates for the random effects. A similar program is written in BMDP (BMDP V5).[116] In addition, S-plus (version 4·0)[117] has procedures for fitting linear and non-linear mixed models. GEE can be fit by commercial software. The SAS GENMOD procedure[115] fits GEE models. The program allows for Gaussian, binary and count data with various link functions (for example, identity, logistic and log-link functions, respectively). The program computes both model-based (under independence, exchangeable, autoregressive and unspecified) and robust estimates of the standard errors for the model parameters. A procedure for fitting GEE in S-plus (version 4·0) is available through STATLAB in the Department of Statistics at Carnegie-Mellon University. A program called SUDAAN[118] also fits GEE models.

Cnaan *et al.*[27] presents SAS code for fitting linear mixed models in their *Statistics in Medicine* tutorial. In the Appendix, I provide the S-plus code used in analysing: (i) the IPPB trial data using Gaussian random-effects models; (ii) the epilepsy clinical trial data using GEE. A discussion of the S-plus output is also provided.

8. DISCUSSION

This paper reviewed statistical methodology for longitudinal data in clinical trials. Many of the most common longitudinal data analysis techniques for continuous and discrete data were reviewed and their application to clinical trials discussed. Recently developed techniques for analysing missing data and for group sequential monitoring of longitudinal data in clinical trials were also discussed.

There is additional methodology which deserves attention. Wei *et al.*[119,120] and Davis *et al.*[121] proposed non-parametric techniques for repeated measures subject to MAR data. Adjusting for compliance is an area of recent work. Rochon[122,123] discussed adjustments for covariates observed post-randomization in linear models for Gaussian data and in a generalized estimating equations framework for discrete and continuous data.

There are many areas of research that need further development. The problem of measurement error in the independent variable in longitudinal model is an example. Analysis in dietary trials often adjust for baseline dietary recall data which is measured with recall error. In addition, these analyses often include treatment-by-time-by-baseline dietary variable interaction terms to assess whether the treatment works differentially with respect to the dietary variable. Ignoring the effect of measurement error can lead to attenuation in assessing these treatment effects. Follmann *et al.*[124] have developed methodology which adjust for measurement error in GEE models with repeated binary responses, and applied this methodology in analysing DISC trial data. Turnbull *et al.*[125] have proposed methodology which adjusts for measurement error in the covariates of regression models for recurrent events. Turnbull *et al.* developed their methodology with a dietary intervention trial similar to the DISC study as a motivating example. Further research is needed to incorporate covariates measured with error into other longitudinal models. The

analysis of discrete longitudinal data with diagnostic error is an area which needs additional work. Espeland *et al.*[126,127] and Albert *et al.*[128] proposed methodology for analysing monotonically increasing binary and ordinal data, respectively, with diagnostic error. These models were applied to longitudinal maturation data similar to what was collected in the DISC study. The development of other models for longitudinal data measured with diagnostics error is an area for further work. The analysis of multiple endpoint data has been an active area for research in methodology for clinical trials.[129,130] Extensions of multiple endpoint methodology to the analysis of longitudinal data is an interesting area for future research. In addition, extensions of the ideas of stochastic curtailment and conditional power to other general classes of longitudinal models such as GEE is an open area for research.

APPENDIX S-PLUS CODE AND OUTPUT

S-plus code is provided for fitting IPPB trial data using a Gaussian random-effects model and for fitting epilepsy clinical trial data using GEE. We begin with the Gaussian random-effects model. The Gaussian random-effects model was fit with the lme routine which is supported by S-plus (version 4·0 for WINDOWS). S-plus code for fitting GEE is not supported by S-plus, but is available for STATLAB at the Department of Statistics, Carnegie-Mellon University. This program can be accessed from the wide world web access:

www.stat.cum.edu/www/cmu-stats/.

The IPPB data set used to fit model (1). The initial data structure was a matrix of 12,545 rows and 4 columns; rows corresponded to observations taken at each time point on each subject, and columns corresponded to subject identity, treatment group, time since randomization, and FEV_1 per cent predicted on the log-scale (the corresponding file was named 'data' and stored on the c drive). The data were entered into the data frame through the command:

```
ippb_read.table ("c:\\data", col.names = c("SUBJECT", "GROUP", "TIM", "FEV")).
```

The procedure is then fit with the command:

```
ippb.fit_lme(fixed = FEV ~ TIM * GROUP,random = ~ TIM,cluster = ~ SUBJECT,data = ippb)
```

where 'fixed' specifies the fixed-effect components of the model, 'random' specifies the random-effect covariates in addition to a random intercept, 'cluster' specifies the variable which denotes the clustering, and 'data' specifies the data frame. The summary outcome can be generated with the command

```
print(summary(ippb.fit))
```

Partial S-plus output is

```
Call:
  Random: ~ TIM
  Fixed: FEV ~ TIM * GROUP
Cluster: ~ SUBJECT
    Data: ippb

Estimation Method: RML
Convergence at iteration: 15
Restricted Loglikelihood: 1649·656
```

```
Restricted AIC:  -3283·312
Restricted BIC:  -3226·855

Variance/Covariance Components Estimate(s):
  Structure: unstructured
  Parametrization: matrixlog
  Standard Deviation (s) of Random Effect(s)
    (Intercept)          TIM
    0·3620197      0·005088593
  Correlation of Random Effects
      (Intercept)
TIM 0·05663255

  Cluster Residual Variance: 0.02420545

Fixed Effects Estimate(s):
              Value Approx.      Std. Error      z ratio(C)
 (Intercept)  -0·0064364816   0·0379724794   -0·1695039
         TIM  -0·0035883561   0·0007734132   -4·6396366
       GROUP  -0·0203453775   0·0241399842   -0·8428082
TIM:GROUP      0·0001390246   0·0004942282    0·2812964


Conditional Correlation(s) of Fixed Effects Estimates

                (Intercept)          TIM          GROUP
         TIM  -0·08996423
       GROUP  -0·94815269     0·08584705
TIM:GROUP      0·08540386    -0·94764509    -0·09078430

Random Effects (Conditional Models):
    (Intercept)          TIM
1    0·2848667206  -7·016515e-003
2    0·1227828924  -3·341595e-003
3   -0·3150682712  -6·386056e-003
4    0·1224302786   2·339991e-003
5    0·0217679391   2·753216e-003
6   -0·3005114682  -7·918168e-004
7   -0·4485412969  -2·436462e-003
8    0·2818809696   3·617564e-004
9   -0·8060045112  -5·719505e-003
10   0·4903508461  -8·733053e-003
11  -0·7936684387   1·667471e-003
12  -0·1569667167   3·507262e-003
13  -0·3952794639  -5·590630e-003
14   0·0809639795   6·445009e-005
15  -0·2664077005  -2·167689e-003
16  -0·4590323703  -6·686447e-004
```

The output provides restricted log-likelihood values to formulate likelihood ratio tests, parameter estimates, standard errors, correlations and z-values for estimates of the fixed effects, along with variance estimates of the random effects. In addition, empirical Bayes estimates of each individual's random effects are provided (only the first 16 are presented here).

The epilepsy clinical trial data was analysed with GEE. The data structure consisted of a data matrix with 295 rows (59 subjects × 5 repeated measurements) with 6 columns. The columns correspond to subject identity, seizure count, duration of each of the observation periods (on the log scale), an indicator of whether an observation period was post-randomization, and a treatment indicator. The data were entered into the data frame with the command

```
epilepsy_read.table("c:\\data",col.names = c("SUBJECT","COUNT",
                    "ti","POST","TREATMENT"))
```

We fit the model

$$\log E(Y_{ij}) = \log t_{ij} + \beta_1 \, \mathrm{POST}_{ij} + \beta_2 \mathrm{TX}_i + \beta_3 \, \mathrm{POST}_{ij} * \mathrm{TX}_i.$$

Once the GEE library is loaded

```
library(gee)
```

the GEE procedure is then fit with the command:

```
ippb.fit_gee (COUNT = FEV ~ offset(ti) + POST * TREATMENT,id = SUBJECT,
              data = epilepsy, family = poisson(link = "log"),
              corstr = "exchangeable"),
```

where 'id' specifies the cluster variable, family specifies the error distribution and link function, and 'corstr' specifies the 'working correlation structure' (other choices are 'independence', 'AR-M', for example). The output resulting from this procedure is:

```
[1] "Beginning Cgee S-function, @(#) geeformula.q 4.13
[1] "running glm to get initial regression estimate"
[1] 1·34760922  0·11183602  0·02753449  -0·10472574
> print (summary (geel.fit))

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4·13 modified 98/01/27 (1998)

Model:
  Link:                      Logarithm
  Variance to Mean Relation: Poisson
  Correlation Structure:     Exchangeable

Call:
gee (formula = COUNT ~ offset (ti) + POST * TREATMENT, id = SUBJECT,
     data = epilepsy, family = poisson (link = "log"), corstr = "exchangeable")

Summary of Residuals:
     Min          1Q      Median      3Q         Max
-4·303571  -1·303571   2·016129  10·37039   147·0444

Coefficients:
                  Estimate   Naive S.E.    Naive z  Robust S.E.   Robust z
(Intercept)     1·34760922   0·1510969   8·9188397   0·1573571   8·5640166
       POST     0·11183602   0·1545145   0·7237900   0·1159304   0·9646821
  TREATMENT     0·02753449   0·2071018   0·1329515   0·2217878   0·1241479
POST: TREATMENT -0·10472579  0·2197052  -0·4766650   0·2134448  -0·4906459

Estimated Scale Parameter:  19·6797
Number of Iterations:  1
```

Working Correlation

	[, 1]	[, 2]	[, 3]	[, 4]	[, 5]
[1,]	1·0000000	0·7713861	0·7713861	0·7713861	0·7713861
[2,]	0·7713861	1·0000000	0·7713861	0·7713861	0·7713861
[3,]	0·7713861	0·7713861	1·0000000	0·7713861	0·7713861
[4,]	0·7713861	0·7713861	0·7713861	1·0000000	0·7713861
[5,]	0·7713861	0·7713861	0·7713861	0·7713861	1·0000000

The output includes parameter estimates, standard-error estimates and z-values obtained assuming the working correlation structure is correct (naive SE and naive z), standard-error estimates and z-values which are robust to misspecification of the correlation structure (robust SE and robust z), and estimates of the working correlation structure. In addition, an estimate of the scale parameter (estimated scale parameter) provides an assessment of over-dispersion.

ACKNOWLEDGEMENTS

I thank Dr. D'Agostino along with the reviewers for comments which improved this tutorial. In addition, I thank Dr. Margaret Wu and Mr. Mario Stylianou for helpful comments on an earlier version of this tutorial.

REFERENCES

1. Diggle, P. J., Liang, K. Y. and Zeger, S. L. *Analysis of Longitudinal Data*, Oxford University Press, Oxford, 1993.
2. Crowder, M. J. and Hand, D. J. *Analysis of Repeated Measures*, Chapman and Hall, London, 1990.
3. Dwyer, J. H., Feinleib, M., Lippert, P. and Hoffmeister, H. *Statistical Models for Longitudinal Studies of Health*, Oxford University Press, Oxford, 1991.
4. Ware, J. H. 'Linear models for the analysis of longitudinal studies', *American Statistician*, **39**, 95–101 (1985).
5. Ashby, 7., Neuhaus, J. M., Hauck, W. W., Pacchetti, P., Heilbron, D. C., Jewell, N. P., Segal, M. R. and Fusaro, R. E. 'An annotated bibliography of methods for analyzing correlated categorical data', *Statistics in Medicine*, **11**, 67–99 (1992).
6. Laird, N. M., Donnelly, C. and Ware, J. H. 'Longitudinal studies with continuous responses', *Statistical Methods in Medical Research*, **1**, 225–247 (1992).
7. Neuhaus, J. M. 'Statistical methods for longitudinal and clustered designs with binary responses', *Statistical Methods in Medical Research*, **1**, 249–273 (1992).
8. Intermittent Positive Pressure Breathing Trial Group. 'Intermittent positive pressure breathing therapy of chronic obstructive pulmonary disease', *Annals of Internal Medicine*, **99**, 612–630 (1983).
9. DISC Collaborative Research Group. 'Dietary Intervention Study in Children (DISC) with elevated low-density-lipoprotein cholesterol. Design and baseline characteristics', *Annals of Epidemiology*, **3**, 393–402 (1993).
10. Kastrukoff, L. F., Orger, J. J., Hashimoto, S. A., Sacks, S. L., Li, D. K., Palmer, M. R., Koopmans, R. A., Petkau, A. J., Berkowitz, J. and Paty, D. W. 'Systemic lymphoblastoid interferon therapy in chronic progressive multiple sclerosis. I. Clinical and MRI evaluation', *Neurology*, **40**, 479–486 (1990).
11. Leppik, I. E. *et al.* 'A double-blind crossover evaluation of progabide in partial seizures', *Neurology*, **35**, 285 (1985).
12. Thall, P. F. and Vail, S. C. 'Some covariance models for longitudinal count data with overdispersion', *Biometrics*, **46**, 657–671 (1990).
13. Johnson, R. E., Jaffe, J. H. and Fudala, P. J. 'A controlled trial of buprenorphine treatment for opiate dependence', *Journal of the American Medical Association*, **267**, 2750–2755 (1992).
14. Pocock, S. J. *Clinical Trials: A Practical Approach*, Wiley, New York, 1983.
15. Morrison, D. F. *Multivariate Statistical Methods*, McGraw Hill, New York, 1976.

16. Rao, C. R. 'Some statistical methods for comparison of growth curves', *Biometrics*, **14**, 1–17 (1958).
17. Rao, C. R. 'Some problems involving linear hypothesis in multivariate analysis', *Biometrika*, **46**, 49–58 (1959).
18. Rao, C. R. 'The theory of least-squares when the parameters are stochastic and its application to the analysis of growth curves', *Biometrika*, **52**, 447–458 (1965).
19. Potthoff, R. and Roy, S. 'A generalized multivariate analysis of variance model useful especially for growth curve problems', *Biometrika*, **51**, 313–326 (1964).
20. Rochon, J. and Helms, R. W. 'Maximum-likelihood estimation for incomplete repeated measures experiments under an ARMA covariance structure', *Biometrics*, **45**, 207–218 (1989).
21. Rochon, J. 'ARMA covariance structures with time heteroscedasticity for repeated measures experiments', *Journal of the American Statistical Association*, **87**, 777–784 (1992).
22. Jones, R. H. *Longitudinal Data With Serial Correlation: A State-Space Approach*, Chapman and Hall, London, 1993.
23. Liang, K. Y. and Zeger, S. L. 'Longitudinal data analysis using generalized linear models', *Biometrika*, **73**, 12–22 (1986).
24. Zeger, S. L. and Liang, K. Y. 'Longitudinal data analysis for discrete and continuous outcomes', *Biometrics*, **42**, 121–130 (1986).
25. McCullagh and Nelder. *Generalized Linear Models.* Chapman and Hall, London, 1989.
26. Laird, N. M. and Ware, J. H. 'Random-effects models for longitudinal data', *Biometrics*, **38**, 963–974 (1982).
27. Cnaan, A., Laird, N. M. and Slasor, P. 'Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data', *Statistics in Medicine*, **16**, 2349–2380 (1997).
28. Lindstrom, M. J. and Bates, D. M. 'Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measures data', *Journal of the American Statistical Association*, **83**, 1014–1022 (1988).
29. Vonesh, E. F. and Carter, R. L. 'Efficient inference for random-coefficient growth curve models with unbalanced data', *Biometrics*, **43**, 617–628 (1987).
30. Waternaux, C., Laird, N. M. and Ware, J. H. 'Methods for analysis of longitudinal data: blood-lead concentrations and cognitive development', *Journal of the American Statistical Association*, **84**, 33–41 (1989).
31. DeGruttola, V., Ware, J. H. and Louis, T. A. 'Influence analysis of generalized least squares estimates', *Journal of the American Statistical Association*, **92**, 911–917 (1987).
32. Weiss, R. E. and Lazaro, C. G. 'Residual plots for repeated measures', *Statistics in Medicine*, **11**, 115–124 (1992).
33. Lange, N. and Ryan, L. 'Assessing normality in random-effects models', *Annals of Statistics*, **17**, 624–642 (1989).
34. Butler, S. M. and Louis, T. A. 'Random effects models with non-parametric priors', *Statistics in Medicine*, **11**, 1981–2000 (1992).
35. Verbeke, G. and Lesaffre, E. 'A linear mixed-effects model with heterogeneity in the random-effects population', *Journal of the American Statistical Association*, **91**, 217–221 (1996).
36. Chi, E. M. and Reinsel, G. C. 'Models for longitudinal data with random-effects and AR(1) errors', *Journal of the American Statistical Association*, **84**, 452–459 (1989).
37. Lindstrom, M. J. and Bates, D. M. 'Nonlinear mixed effects models for repeated measures data', *Biometrics*, **46**, 673–687 (1990).
38. Vonesh, E. F. and Carter, R. L. 'Mixed-effects nonlinear regression for unbalanced repeated measures', *Biometrics*, **48**, 1–17 (1992).
39. Davidian, M. and Giltinan, D. M. *Nonlinear Models For Repeated Measures*, Chapman and Hall, London, 1995.
40. Morrell, C. H., Person, J. D., Ballentine, H., Carter, R. L. and Brant, L. J. 'Estimating unknown transition times using a piecewise nonlinear mixed-effects model in men with prostate cancer', *Journal of the American Statistical Association*, **90**, 45–53 (1995).
41. Vonesh, E. F., Chinchilli, V. M. and Pu, K. 'Goodness-of-fit in generalized nonlinear mixed-effects models', *Biometrics*, **52**, 572–587 (1996).
42. Fitzmaurice, G. M. 'A caveat concerning independence estimating equations with multivariate binary data', *Biometrics*, **51**, 309–317 (1995).

43. Albert, P. S. and McShane, L. M. 'A generalized estimating equations approach for spatially correlated binary data: applications to the analysis of neuroimaging data', *Biometrics*, **51**, 627–638 (1995).
44. Rotnitzky, A. and Jewell, N. P. 'Hypothesis testing of regression parameters in semi-parametric generalized linear models for clustered correlated data', *Biometrika*, **77**, 485–497 (1990).
45. Prentice, R. L. 'Correlated binary regression with covariates specific to each binary observation', *Biometrics*, **44**, 1033–1048 (1988).
46. Zhao, L. P. and Prentice, R. L. 'Correlated binary regression using a generalized quadratic model', *Biometrika*, **77**, 642–648 (1990).
47. Liang, K. Y., Zeger, S. L. and Qaqish, B. 'Multivariate regression analyses for categorical data (with discussion)', *Journal of the Royal Statistical Society, Series B*, **54**, 3–40 (1992).
48. Podgor, M. J., Hiller, R. and The Framingham Eye studies Group. 'Associations of types of lens opacities between and within eyes of individuals: An application of second-order generalized estimating equations', *Statistics in Medicine*, **15**, 145–156 (1996).
49. Miller, M. E., Davis, C. S. and Landis, J. R. 'The analysis of longitudinal polytomous data: generalized estimating equations and connections with weighted least squares', *Biometrics*, **49**, 1033–1044 (1993).
50. Lipsitz, S. R., Kim, K. and Zhao, L. 'Analysis of repeated categorical data using generalized estimating equations', *Statistics in Medicine*, **13**, 1149–1163 (1994).
51. Heagerty, P. J. and Zeger, S. L. 'Marginal regression models for clustered ordinal measurements', *Journal of the American Statistical Association*, **91**, 1024–1035 (1996).
52. Liu, G. and Liang, K. Y. 'Sample size calculations for studies with correlated observations', *Biometrics*, **53**, 937–947 (1997).
53. Preisser, J. S. and Qaqish, B. F. 'Deletion diagnostics for generalized estimating equations', *Biometrika*, **83**, 551–562 (1996).
54. Heagerty, P. J. and Zeger, S. L. 'Lorelogram: A regression approach to exploring dependence in longitudinal categorical responses', *Journal of the American Statistical Association*, **93**, 150–162 (1998).
55. Barnhart, H. X. and Williamson, J. M. 'Goodness-of-fit tests for GEE modeling with binary responses', *Biometrics*, **54**, 326–335 (1998).
56. Fitzmaurice, G. M. and Laird, N. M. 'A likelihood-based method for analysing longitudinal binary data', *Biometrika*, **80**, 141–151 (1993).
57. Lang, J. and Agresti, A. 'Simultaneously modeling joint and marginal distributions of multivariate categorical responses', *Journal of the American Statistical Association*, **89**, 625–632 (1994).
58. Molenberghs, G. and Lesaffre, E. 'Marginal modeling of correlated ordinal data using an n-way Plackett distribution', *Journal of the American Statistical Association*, **89**, 633–644 (1994).
59. Lindsey, J. K. and Lambert, P. 'On the appropriateness of marginal models for repeated measurements in clinical trials', *Statistics in Medicine*, **17**, 447–469 (1998).
60. Zeger, S. L., Liang, K. Y. and Albert, P. S. 'Models for longitudinal data: a generalized estimating equation approach', *Biometrics*, **44**, 1049–1060 (1988).
61. Neuhaus, J. M., Kalbfleisch, J. D. and Hauck, W. W. 'A comparison of cluster-specific and population-averaged approaches for analysing correlated binary data', *International Statistical Review*, **59**, 25–35 (1990).
62. Neuhaus, J. M. 'Estimation efficiency and tests of covariate effects with clustered data', *Biometrics*, **49**, 989–996 (1993).
63. Stiratelli, R., Laird, N. and Ware, J. H. 'Random effects models for serial observations with binary responses', *Biometrics*, **40**, 961–971 (1984).
64. Longford, N. T. *Random Coefficient Models*, Oxford University Press, Oxford, 1993.
65. Waclawiw, M. A. and Liang, K. Y. 'Prediction of random effects in the generalized linear model', *Journal of the American Statistical Association*, **88**, 171–178 (1993).
66. Zeger, S. L. and Karim, M. R. 'Generalized linear models with random effects: A Gibbs' sampling approach', *Journal of the American Statistical Association*, **86**, 79–95 (1991).
67. Breslow, N. E. and Clayton, D. G. 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association*, **88**, 9–25 (1993).
68. Follmann, D. A. and Lambert, D. 'Generalized logistic regression by non-parametric mixing', *Journal of the American Statistical Association*, **84**, 295–300 (1989).
69. Cox, D. R. *The Analysis of Binary Data*, Chapman and Hall, London, 1970.

70. Muenz, L. R. and Rubinstein, L. V. 'Markov models for covariate dependence of binary sequence', *Biometrics*, **43**, 863–871 (1985).
71. Zeger, S. L. and Qaqish, B. 'Markov regression models for time series: A quasi-likelihood approach', *Biometrics*, **44**, 1019–1031 (1988).
72. Albert, P. S. 'A transitional model for longitudinal binary data subject to nonignorable missing data', submitted for publication (1999).
73. Fahrmeir, L. and Kaufmann, H. 'Regression models for nonstationary categorical time series', *Journal of Time Series Analysis*, **8**, 147–160 (1987).
74. Albert, P. S. and Brown C. H. 'The design of a panel study under an alternating Poisson assumption', *Biometrics*, **47**, 921–932 (1991).
75. Kalbfleish, J. D. and Lawless, J. F. 'The analysis of panel data under a Markov assumption', *Journal of the American Statistical Association*, **80**, 863–873 (1985).
76. Kosorok, M. R. and Chao, W. H. 'The analysis of longitudinal ordinal response data in continuous time', *Journal of the American Statistical Association*, **91**, 807–817 (1996).
77. Cook, R. J. and Ng, E. T. M. 'A logistic-bivariate normal model for overdispersed two-state Markov processes', *Biometrics*, **53**, 358–364 (1997).
78. Albert, P. S. and Waclawiw, M. A. 'A two-state Markov chain for heterogeneous transitional data: a quasi-likelihood approach', *Statistics in Medicine*, **17**, 1481–1493 (1998).
79. Lawless, J. F. 'Regression methods for Poisson process data', *Journal of the Americal Statistical Association*, **82**, 808–815 (1987).
80. Thall, P. F. 'Mixed Poisson likelihood regression models for longitudinal interval count data', *Biometrics*, **44**, 197–209 (1988).
81. Abu-Libdeh, H., Turnbull, B. W. and Clark, L. C. 'Analysis of multi-type recurrent events in longitudinal studies; application to a skin cancer prevention trial', *Biometrics*, **46**, 1017–1034 (1990).
82. Wei, L. J., Lin, D. Y. and Weissfeld, L. 'Regression analysis of multivariate incomplete failure time data by modeling marginal distributions', *Journal of the American Statistical Association*, **84**, 1065–1073 (1989).
83. Cook, R. J., Lawless, J. F. and Nadeau, C. 'Robust tests for treatment comparisons based on recurrent event responses', *Biometrics*, **52**, 557–571 (1996).
84. Cook, R. J. 'The design and analysis of randomized trials with recurrent events', *Statistics in Medicine*, **14**, 2081–2098 (1995).
85. Barnhart, H. X. and Sampson, A. R. 'Multiple population models for multivariate random length data-with applications in clinical trials', *Biometrics*, **51**, 195–204 (1995).
86. Albert, P. S., Follmann, D. A. and Barnhart, H. X. 'A generalized estimating equation approach for modeling random length binary vector data', *Biometrics*, **53**, 376–384 (1997).
87. Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data*, Wiley, New York, 1987.
88. Rotnitzky, A. and Wypij, D. 'A note on the bias of estimators with missing data', *Biometrics*, **50**, 1163–1170 (1994).
89. Robins, J. M., Rotnitzky, A. and Zhao, L. P. 'Analysis of semi-parametric regression models for repeated outcomes in the resence of missing data', *Journal of the American Statistical Association*, **90**, 106–121 (1995).
90. Wu, M. C. and Carroll, R. J. 'Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process', *Biometrics*, **44**, 175–188 (1988).
91. Wu, M. C. and Bailey, K. R. 'Estimation and comparison of changes in the presence of informative right censoring: conditional linear model', *Biometrics*, **45**, 939–955 (1989).
92. Mori, M., Woolson, R. F. and Woodworth, G. G. 'Slope estimation in the presence of informative right censoring: modeling the number of observations as a geometric random variable', *Biometrics*, **50**, 39–50 (1994).
93. Follmann, D. A. and Wu, M. C. 'An approximate generalized linear model with random effects for informative missing data', *Biometrics*, **51**, 151–168 (1995).
94. Stasny, E. A. 'Some Markov-chain models for nonresponse in estimating gross labor force flows', *Journal of Official Statistics*, **3**, 359–373 (1987).
95. Conaway, M. R. 'Non-ignorable non-response models for time-ordered categorical variables', *Applied Statistics*, **42**, 105–115 (1993).

96. Diggle, P. J. and Kenward, M. G. 'Informative dropout in longitudinal data analysis (with discussion)', *Applied Statistics*, **43**, 49–93 (1994).
97. Molenberghs, G., Kenward, M. G. and Lesaffre, E. 'The analysis of longitudinal ordinal data with nonrandom drop-out', *Biometrika*, **84**, 33–44 (1997).
98. Cook, R. J. and Lawless, J. F. 'Marginal analysis of recurrent events and a terminating event', *Statistics in Medicine*, **16**, 911–924 (1997).
99. Dawson, T. D. and Lagakos, S. W. 'Size and power of two sample tests of repeated measure', *Biometrics*, **49**, 1022–1032 (1993).
100. Follmann, D., Wu, M. and Geller, N. L. 'Testing treatment efficacy in clinical trials with repeated binary measurements and missing observations', *Communications in Statistics – Theory and Methods*, **23**, 557–574 (1994).
101. Pocock, S. J. 'Group sequential methods in the design and analysis of clinical trials', *Biometrika*, **64**, 191–199 (1977).
102. O'Brien, P. C. and Fleming, T. R. 'A multiple testing procedure for clinical trials', *Biometrics*, **35**, 549–556 (1979).
103. Slud, E. and Wei, 'Two-sample repeated significance tests based on the modified Wilcoxon statistics', *Journal of the American Statistical Association*, **77**, 862–868 (1982).
104. Lan, K. K. G. and DeMets, D. L. 'Discrete sequential boundaries for clinical trials', *Biometrika*, **70**, 659–663 (1983).
105. Geary, D. N. 'Sequential testing in clinical trials with repeated measurements', *Biometrika*, **75**, 311–318 (1988).
106. Lee, J. W. and DeMets, D. L. 'Sequential comparison of changes with repeated measurements data', *Journal of the American Statistical Association*, **86**, 757–762 (1991).
107. Wu, M. C. and Lan, K. K. 'Sequential monitoring for comparison of changes in a response variable in clinical studies', *Biometrics*, **48**, 765–779 (1992).
108. Wei, L. J., Su, J. Q. and Lachin, J. M. 'Interim analyses with repeated measurements in a sequential clinical trial', *Biometrika*, **77**, 359–364 (1990).
109. Gange, S. J. and DeMets, D. L. 'Sequential monitoring of clinical trials with correlated responses', *Biometrika*, **83**, 157–167 (1996).
110. Cook, R. J. and Lawless, J. F. 'Interim monitoring of longitudinal comparative studies with recurrent event response', *Biometrics*, **52**, 1311–1323 (1996).
111. Lan, K. K. G., DeMets, D. L. and Halperin, M. 'More flexible sequential and nonsequential designs in long-term clinical trials', *Communications in Statistics*, **A13**, 2339–2353 (1984).
112. McMahon, R. P., Waclawiw, M. A., Geller, N. L., Baron, F. B., Terrin, M. C. and Bonds, D. R. 'An extension of stochastic curtailment for incompletely reported and classified recurrent events: the multi center study of hydroxyurea in sickle cell anemia MSH)', *Controlled Clinical Trials*, **18**, 420–430 (1997).
113. Lan, K. K. G. and Zucker, D. M. 'Sequential monitoring of clinical trials: the role of information and brownian motion, *Statistics in Medicine*, **12**, 753–766 (1993).
114. Halperin, M., Lan, K. K. G., Wright, E. C. and Foulkes, M. A. 'Stochastic curtailing for comparison of slopes in longitudinal studies', *Controlled Clinical Trials*, **8**, 315–326 (1987).
115. SAS Institute Inc., *SAS/STAT Users Guide, Version 6, fourth edition, volume, 2*, Cary, NC.
116. BMDP, *Statistical Software Manual.* University of California Press, Berkeley.
117. MathSoft Inc., *S-PLUS version 4·0, User's Guide*, MathSoft, Inc. Seattle, Washington.
118. Bieler, G. S. and Williams, R. L. *Application of the SUDAAN Software Package to Clustered Data Problems*, Research Triangle Institute, Research Triangle Park, NC, 1996.
119. Wei, L. J. and Lachin, J. M. 'Two sample asymptotically distributionfree tests for incomplete multivatiate observations', *Journal of the American Statistical Association*, **79**, 653–661 (1984).
120. Wei, L. J. and Johnson, W. E. 'Combining dependent tests with incomplete repeated measurements', *Biometrika*, **72**, 359–364 (1985).
121. Davis, C. S. and Wei, L. J. 'Nonparametric method for analyzing incomplete nondecreasing repeated measurements', *Biometrics*, **44**, 1005–1018 (1988).
122. Rochon, J. 'Supplementing the intent-to-treat analysis: accounting for covariates observed post randomization in clinical trials', *Journal of the American Statical Association*, **90**, 292–300 (1995).
123. Rochon, J. 'Accounting for covariates observed post randomization for discrete and continuous repeated measures data', *Journal of the Royal Statistical Society, Series B*, **58**, 205–219 (1996).

124. Follmann, D A., Hunsberger, S. A. and Albert, P. S. 'Repeated probit regression when covariates are measured with error', *Biometrics*, In press, 1999.
125. Turnbull, B. W., Jiang, W. and Clark, L. C. 'Regression models for recurrent event data: parametric random-effects models with measurement error', *Statistics in Medicine*, **16**, 853–864 (1997).
126. Espeland, M. A., Platt, O. S. and Gallagher, D. 'Joint estimation of incidence and error rates from irregular longitudinal data', *Journal of the American Statistical Association*, **84**, 972–979 (1989).
127. Espeland, M. A., Rushing, J. T. and DeVault, A. 'Estimating incidence and diagnostic error rates for bivariate progressive processes', *Biometrics*, **49**, 1010–1021 (1993).
128. Albert, P. S., Hunsberger, S. A. and Biro, F. M. 'Modeling repeated measures with monotonic ordinal responses and misclassification, with applications to studying maturation', *Journal of the American Statistical Association*, **92**, 1304–1312 (1997).
129. Pocock, S. J., Geller, N. L. and Tsiatis, A. A. 'The analysis of multiple endpoint data in clinical trials', *Biometrics*, **43**, 487–498 (1987).
130. Lefkopoulou, M. and Ryan, L. 'Global tests for multiple binary outcomes', *Biometrics*, **49**, 975–988 (1993).

TUTORIAL IN BIOSTATISTICS

Repeated measures in clinical trials: simple strategies for analysis using summary measures

Stephen Senn*,†, Lynda Stevens and Nish Chaturvedi

Department of Epidemiology and Public Health, University College London, 1-19 Torrington Place, London WC1E 6BT, U.K.

SUMMARY

The summary measures approach to analysing repeated measures is described. The circumstances under which it can be advantageous to use such measures are considered. Strategies for baseline adjustment where there are multiple baselines are examined, as is the choice of appropriate summary statistic. A compromise trend/mean measure, regression through the origin, is proposed as being useful under some circumstances. An analysis using this measure is illustrated with a suitable example. Copyright © 2000 John Wiley & Sons, Ltd.

INTRODUCTION

Clinical trials in which there is a single endpoint are important but unusual. In such trials it is almost always the case that an acute disease is being studied and that either time until death or time until cure is the single endpoint. For trials in chronic diseases, in which it is expected that the patient will still be alive and still suffering from the disease by the end of the period of study, it is common to take repeated measurements during the course of the trial. Usually a number of variables are measured on a number of occasions both before (baseline) and after commencing therapy (outcome). Where the measurements made are continuous, a popular, simple and effective approach to analysing such trials is the so-called 'summary measures' approach [1–3] (also known as the 'derived variables' approach). In this paper we consider some issues which affect the implementation of the summary measures approach.

The summary measures approach consists of these two stages. First, a summary descriptive statistic is calculated for each patient using the set of repeated measures. (Usually the summary descriptive statistic is calculated from repeated measures on the same variable at different time points rather than, say, different variables at the same time point, although O'Brien has suggested

* Correspondence to: Stephen Senn, Department of Epidemiology and Public Health, University College London, 1-19 Torrington Place, London WC1E 6BT, U.K.
† E-mail: stephens@public-health.ucl.ac.uk

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
 ISBN: 0-470-02365-1

an approach whereby different variables can be combined [4]. Unless explicitly stated otherwise, we shall consider that in what follows summaries over time of the same variable are involved.) Second, formal modelling begins once the summary measures have been constructed and these summary measures are then used as endpoints in a more conventional analysis.

For example, in a three-month placebo-controlled trial of a bronchodilator in asthma, patients might take readings of peak expiratory flow on waking every morning. A suitable summary measure might be the mean of these measures over the three months. These means could then be used to compare the two groups. In a two-year placebo-controlled trial of the effect of hormone replacement therapy on osteoporosis, patients might have their bone mineral density measured every six months during a visit to the clinic. A popular choice of summary measure might then be the estimated slope of the bone mineral density over the two years. Again these slopes could then be used to compare the two groups.

The issues which we consider below are the following.

1. When is the summary measures approach appropriate?
2. What is the role of baseline measurements?
3. What should one do when more than one baseline measurement is available?
4. Under what circumstances should one use means and under what circumstances should one use slopes?
5. Are compromise measures possible?

However, before considering these issues in connection with the summary measures approach, we briefly consider some popular alternatives.

ALTERNATIVES TO THE SUMMARY MEASURES APPROACH

Finney has provided a useful review of approaches to repeated measures [5]. These include the following:

1. *Multivariate analysis.* Here the repeated measures over time are analysed using multivariate analysis of variance (MANOVA) or some equivalent technique. For example, for the case with two treatment groups, the repeated measures could be compared using Hotelling's T^2 statistic. This approach makes no use of the sequential ordering of the measurements through time, since permuting the time points will give exactly the same answer. Such a permutation, however, would make the pattern of such results quite different and the invariance which this analysis shows to such permutations must be seen as a weakness. Also, the question that the approach answers, 'do the treatments differ in any way over time?' is too general to be of much practical interest. The approach does not yield useful estimates of the effect of treatment.
2. *Split-plot analysis.* Here time is treated as a factor. The mean level of the treatments is compared and (more controversially) the time by treatment interaction is examined. This approach has remained popular with psychologists but it has a number of inherent limitations that have made it unpopular with medical statisticians. Testing the time by treatment interaction either involves the strong assumption of sphericity of the data or complex and generally conservative adjustments. Effectively, an implicit assumption in the approach is that although patients may vary as regards their mean levels, they do not vary

as regards response to treatment. As regards interpretation, it has all the disadvantages of the multivariate approach and shows the same permutation invariance.

3. *Independent analysis.* Here an analysis is made of outcomes at each time point ignoring the data from other time points. An advantage of this approach is its extreme simplicity. The disadvantage is that it is inefficient and does not reduce the data effectively. The approach reduces the computational difficulties in repeated measures but does not bring a corresponding reduction in interpretation. If (as will surely be the case) repeated measures are correlated, there is a danger that evidence which is not independent will be over-interpreted. On the other hand, the fact that measures will not be perfectly correlated means that the probability that at least one measure will be significant is increased. However, this approach can be attractive when the trial has very high precision and can also be useful under circumstances for a preliminary analysis as a means of identifying an appropriate model. (This strategy, however, is not acceptable in a regulatory context where prespecified analyses are required.)
4. *Random coefficients.* This approach goes by a number of names, including, for example, random growth curve models, multi-level models, random effect models (of which it is a special case), hierarchical models. The response over time shown by a particular patient is viewed as being determined by a few parameters (generally fewer than the number of time points [6,7]. These parameters are themselves regarded as being sampled from some hyper-distribution. In more sophisticated approaches a model is also included for time dependent autocorrelation within patients [8]. The advantages of the approach include that it takes explicit account of posited dependencies between data, yields useful global estimates and permits shrunk individual estimates to be produced and even individual predictions. Disadvantages include the degree of modelling that is necessary, the fact that many different choices of model are possible and that model fitting may be difficult. With a number of packages now available, the latter is less of a problem than it was until recently.

These are some of the most popular approaches used for analysing continuous data. For non-linear models there are other distinctions which are important but they do not concern us here. Many medical statisticians now agree that approach 2 is illegitimate and approach 1 of little practical use. Approach 3 can be useful on occasion if the number of repeated measures per patient is not great. Approach 4 is probably the most useful and it is as an alternative to this that the summary approach measures can often be employed. We now consider this point.

WHEN IS THE SUMMARY MEASURES APPROACH USEFUL?

A summary measures analysis which, as regards estimation of the treatment effects, corresponds to the random coefficients model, can usually be produced, provided that data are not sparse. If complete data are available and measurements are taken at the same time points for all patients, then the summary measures approach is often completely efficient. A rather unusual concrete example may help to make the point.

Consider a cross-over design to compare two treatments, A and B, in four periods and two sequences. Suppose for the moment, that patients have been allocated at random to the sequences ABAB and BABA. Carry-over is assumed not to occur. The object is to calculate the treatment contrast: the difference between the effect of B and the effect of A. Various models could be

Table I. Possible variance–covariance structure (extract) for a cross-over trial

		Patient 1				Patient 2			
		1	2	3	4	1	2	3	4
Patient 1	1	σ^2	$\rho\sigma^2$	$\rho\sigma^2$	$\rho\sigma^2$	0	0	0	0
	2		σ^2	$\rho\sigma^2$	$\rho\sigma^2$	0	0	0	0
	3			σ^2	$\rho\sigma^2$	0	0	0	0
	4				σ^2	0	0	0	0
Patient 2	1					σ^2	$\rho\sigma^2$	$\rho\sigma^2$	$\rho\sigma^2$
	2						σ^2	$\rho\sigma^2$	$\rho\sigma^2$
	3							σ^2	$\rho\sigma^2$
	4								σ^2

entertained for this cross-over trial. For example

$$Y_{i,j} = \mu + \pi_j + \tau_{(i,j)} + \varepsilon_{i,j} \tag{1}$$

where $Y_{i,j}$ is the response observed in period j for patient i, π_j is an effect due to period j, $\tau_{(i,j)}$ is the effect of the treatment applied to patient i in period j and $\varepsilon_{i,j}$ is a random disturbance term with expectation zero. Various assumptions can be made regarding the $\tau_{(i,j)}$ and $\varepsilon_{i,j}$ terms. The $\varepsilon_{i,j}$ are often assumed to have a 'block diagonal' form with $\text{cov}(\varepsilon_{i,j}, \varepsilon_{i^*,j^*})$ equal to σ^2 if $i = i^*$ and $j = j^*$, equal to $\rho\sigma^2$ if $i = i^*$ and $j \neq j^*$ and equal to 0 otherwise. Such a block diagonal form is illustrated in Table I for the first two patients in such a clinical trial. In plain English what it assumes is independence (and hence zero covariance and correlation) between measurements made on different patients, a constant variance σ^2 for all measurements made on the same patients and constant correlation, ρ, and hence covariance $\rho\sigma^2$, for different measurements on the same patient. (This 'block diagonal' form arises if we consider that there are two independent sources of variation: between and within patient. It is an alternative to using a model with a fixed patient effect.) The term $\tau_{(i,j)}$ is often taken as a fixed effect which depends only on the treatment given to patient i in period j but does not vary according to the patient. (This is not always wise. It may be useful to allow that the treatment effect may vary from patient to patient.)

Under such circumstances, an analysis which calculates for each patient the mean difference between the results under treatment B and treatment A (a summary measure), averages these within each sequence and then across both sequences gives the same estimate as fitting generalized least squares to the whole data. Such a summary measure would use the weights -0.25, 0.25, -0.25, 0.25 for the four readings for patients from the first sequence and 0.25, -0.25, 0.25, -0.25 for patients from the second sequence.

If, however, we allowed that the correlation between repeated measures on the same patient had the more general form

$$\text{corr}(Y_{ij}, Y_{ij^*}) = \rho_1 + \rho_2^{|j-j^*|}(1 - \rho_1) \tag{2}$$

a model which corresponds to there being an autoregressive process for the within-patient disturbance terms, then the summary measures approach is no longer efficient, unless $\rho_2 = 0$, in which case the correlation structure reduces to that previously considered. In general, the optimal weights depend on the value of ρ_2. For example, if this is 0.5, then the optimal scheme of weights is -0.2, 0.3, -0.3, 0.2 and 0.2, -0.3, 0.3, -0.2. Also, suppose that there are a number of

patients who fail to complete the trial for reasons unconnected with the efficacy of the treatment. If they have completed three or even two periods we shall be able to calculate a treatment difference for such patients. However, being based on fewer observations we ought to give such summaries lesser weight, but nothing in the simple summary measures approach enables us to do so. Thus, the approach is inefficient in such circumstances. Similarly, suppose we had chosen the ABBB/BAAA design instead. This is unbalanced in some sense since the numbers of periods allocated to the two treatments are not identical for all patients. The obvious weights for the summary statistic are $-1/2$, $1/6$, $1/6$, $1/6$ and $1/2$, $-1/6$, $-1/6$, $-1/6$. However, if we accept that 'patient' is a random effect, that is to say that the general level of outcome that would be seen in the absence of treatment, varies randomly from patient to patient, as is always implicitly assumed in a parallel group trial, these weights are not fully efficient even if the more simple of the two correlation structures applies.

These three cases – a more complex correlation structure, missing data, a design which is unbalanced in some sense – are cases where a *simple* summary measures approach will not be fully efficient. However, it does not follow that a summary measures analysis is not appropriate. Consider the case where the correlation structure in (2) is deemed to apply. Either (which is unlikely), the value of ρ_2 will be known, in which case a more efficient set of weights for the summary measure *could* be constructed, or it will not, in which case the value of ρ_2 will have to be estimated in any alternative analysis. This brings its own dangers, not least that the estimates of the variance of the treatment effect will be biased, and, of course, the presumed model for the correlation structure may in any case be quite wrong.

This is one of the advantages of the summary measures approach. The issue of the correlation between measures is fitnessed, rendered irrelevant by the tactic employed. No estimate of this structure is required. It is only the joint effect of variances and covariances on the variability of the measure itself that matters and this is estimated directly from the variance of the measure between patients. This empirical device eliminates the need for theorizing about correlation structure.

Since missing data are often encountered in clinical trials and complicated correlation structures may frequently be supposed to apply, a summary measures approach will rarely be optimal. In practice, however, it will often be very good. There are exceptions. Consider, as an analogy, a multi-centre trial with (as is usually the case) a very uneven distribution of patients per centre. Suppose that we wish to regard the effect of a treatment in a particular centre as random. We can then regard the patients as repeated measures on the centre. If, however, we use the summary measures approach on the *centres*, first, we shall estimate a treatment effect for each centre, and second, weight them all equally, ignoring the fact that very different numbers of patients have been measured. This will lead to a considerable loss of efficiency in estimation and most statisticians would regard this as too extreme in such a case.

It could also be the case that we can interested in the correlation structure of the data. In the case of the cross-over trial above, we might wish to explore the extent to which the effect of treatment varied from patient to patient: the patient by treatment interaction. This requires separating the random effect from the true within-patient error and this cannot be done using summary measures which jointly incorporates these two sources of variability. In fact, the other side of the coin to the robustness of the summary measures approach to assumptions regarding variance and covariance structure is its inability to examine that structure.

However, the example is a little unusual. Multi-period cross-over trials are, in fact, particularly well suited to estimating patient by treatment interaction because the patients are given different

treatments in a number of periods. Conventional repeated measures designs, in which the patient is given one treatment over a long period, require an assumption that the response over time can be modelled using fewer parameters than there are observations if pure within-patient error is to be separated from the random effect.

Thus, to sum up, if we are interested in patient by treatment interaction then the random coefficients approach should be used. However, provided we are primarily interested in the mean effect of treatment, the summary measures approach will often be very useful provided the information available per patient does not vary too much. It is difficult to be precise as to just how much variation is too much. Consider the case where we are producing a mean of summary measures calculated from N patients and the variance of the measure on patient i is v_i and the random effect variance is v_b and let the total variance for a given patient be $V_i = v_i + v_b$. The mean of the N summaries will have variance

$$V_{\text{sum}} = \frac{1}{N^2} \sum_{i=1}^{N} V_i = \frac{1}{N} \bar{V}_{\text{a}}$$

where $\bar{V}_{\text{a}}$ is the arithmetic mean of the total variances for each patient. Now, if the v_i are identical for every patient the optimal weighting is indeed equal for each summary measure, but if not, weights proportional to the $W_i = 1/V_i$ will be optimal. The resulting estimator has variance

$$\bar{V}_{\text{opt}} = \frac{\sum_{i=1}^{N} W_i^2 V_i}{(\sum_{i=1}^{N} W_i)^2} = \frac{1}{\sum_{i=1}^{N} W_i} = \frac{1}{N} \bar{V}_{\text{h}}$$

where $\bar{V}_{\text{h}}$ is the harmonic mean of the total variances for each patient. Unless all values are identical, the harmonic mean is always less than the arithmetic mean so that V_{opt} is, as it should be, less than or equal to V_{sum}. The loss in applying the summary measures approach is thus given by the extent to which the arthmetic mean of the variances, including the common v_b, exceeds the harmonic mean of these variances.

WHAT IS THE ROLE OF BASELINE MEASUREMENT?

In our view it is not useful to use baselines to *define* effects. They may be very useful in helping to *measure* effects. That is to say, one should not naively regard the effect of treatment as being the difference between outcome and baseline but baselines may none the less be useful in measuring the effect. If one defines the effect of treatment as being the difference between what happened as a result of treatment and what would have happened had treatment been denied (or had an alternative treatment been given), then since by definition the baseline cannot measure the result of treatment, its role, if any, must be in predicting what would have happened.

All too often in the medical literature the treatment response at time t is *defined* first for a patient i as $d_i = Y_{it} - Y_{i0}$, where Y_{it} is the measurement taken at time t and Y_{i0} is the baseline reading. These 'responses' (sometimes referred to as the 'change scores') are then used as outcome measures in a conventional analysis. As a *definition* of response this only makes sense if the baseline reading is regarded as predicting what would have happened to the patient had treatment been denied. The definition is thus naive and runs contrary to the spirit of the controlled clinical trial. However, the further analysis of the change scores restores the role of the control and what emerges as a result is an unbiased estimate of the treatment effect.

From one point view, the analysis of change-scores is a perfectly reasonable application of the summary measures approach. Two measurements are reduced to one in a first step and the resulting statistics are then analysed. An alternative approach would be to ignore the baseline value altogether and simply use the raw outcomes, Y_{it}. As is well known, if the variance of baseline and outcome are equal, then the change-score approach is superior provided that the *partial* correlation of outcome and baseline (given the factors allowed for in the model for analysis) is greater than 0.5. This is commonly the case in parallel group trials provided only the treatment effect is allowed for in the model but is often not the case in cross-over trials, where patient and perhaps period effects will also be included.

As is also well known, if the baseline is included in the model as a covariate, then the change score and raw outcomes approaches are equivalent [9]. Indeed, change scores and raw outcomes can be regarded as special cases of summary statistics of the form $S_{it} = (Y_{it} - \delta Y_{i0})$, where the change score results when $\delta = 1$ and the raw outcome results when $\delta = 0$. Analysis of covariance uses a value for δ which minimizes the variance of S_{it} or (equivalently) makes S_{it} independent of the baseline. These are strong arguments in favour of using analysis of covariance.

The only disadvantage of analysis of covariance is that the slope coefficient δ has to be estimated. This brings with it a slight penalty in terms of efficiency as, inevitably, the baselines will be perfectly balanced between the two groups (Reference [10], chapter 7).

WHAT SHOULD ONE DO WHEN MORE THAN ONE BASELINE MEASUREMENT IS AVAILABLE?

In many clinical trials more than one baseline measurement will be available. For example, it is not uncommon for clinical trials to have a run-in period prior to randomization. It is then usual to have at least two baseline measurements available for patients: one when first selected for entry into the trial and another just prior to randomization. There may also have been intervening measurements. Given that it has been agreed that these baseline measurements should be exploited in an analysis of covariance, there would then seem to be three obvious choices for using them. First, one could simply use the last of all these baseline measurements. Second, one could use the mean of all the baseline measurements. Third, one could fit all of the baselines simultaneously as covariates. In the paragraphs which follow we give some intuitively reasonable guidance regarding choice of strategy. Further discussion is given in *Statistical Issues in Drug Development* [10], chapter 7.

An argument for using the latest baseline measurement only in analysis of covariance is that it is likely to be the most strongly predictive of all the measurements. Thus if a choice has to be made of one baseline only, then this would seem to be the best choice. In practice, however, this will not be fully efficient unless the partial correlation between outcome and previous baselines is zero given the most recent baseline [10]. This corresponds to an autoregressive process of order 1. That is to say, the correlation between successive measurements is of the form ϕ, ϕ^2, ϕ^3 etc. An intuitive explanation is that analysis of covariance using baselines helps to eliminate (to an appropriate degree) the difference between patients. Because of measurement error, a single baseline measurement will not be sufficient to establish the true condition of a patient at baseline. Hence, further control for this source of variation could be achieved by using more measurements.

Given that the variances of the baseline measurements are equal, then using the mean of the baseline measurements turns out to be a reasonable strategy if the correlations of the baselines with the outcome measurement are the same. Where this is so, the baselines are exchangeable for the purpose of prediction and nothing is to be gained by fitting them separately. If, however, there were appreciable differences between the various correlations between baselines and outcomes then there might be some value in fitting the baselines separately. For example, it might be the case that patients were subject to differing trends. That being so, simply using the mean baseline would not be as effective at predicting the value at outcome as using (say) the mean and a slope estimate based on the baselines. Equivalently, the two baselines could be fitted.

WHEN CHOOSING SUMMARY MEASURES, UNDER WHAT CIRCUMSTANCES SHOULD ONE USE MEANS AND UNDER WHAT CIRCUMSTANCES SHOULD ONE USE SLOPES?

The simple answer to this question is that the means are useful if the effect of treatment is expected to be almost immediate and will then be maintained steadily, whereas slopes are more appropriate if the effect of treatment is expected to grow at a constant rate over time. These two cases are illustrated in Figure 1, which shows, in the left hand panel, the supposed situation for forced expiratory volume in one second (FEV_1) for a placebo-controlled trial in asthma and in the right hand panel the supposed position for bone mineral density for a trial in osteoporosis. It should be noted that in practice it is the second of the two cases that corresponds to the more simple function for the effect of treatment over time. The first assumes that there is some sort of change in the function; a period when the effect of treatment grows very rapidly and then levels off. In many trials in which means are used as summary statistics, it is implicitly assumed that all measurements of outcome are made after the period of growth of effect has ended.

Consider a case where measurements are taken on n_h patients in treatment group h, $h = 1, 2$ at times $t_j, j = 0, 1, \ldots, k$ where t_0 is the time at baseline. We may represent the general situation using the model

$$Y_{h(i)j} = \mu + \pi(t_j) + \tau_h(t_j) + \phi_{h(i)} + \gamma_{h(i)}(t_j) + \varepsilon_{h(i)j} \tag{3}$$

Here, $Y_{h(i)j}$ is the measurement on subject i of treatment group h in period j, $\pi(t_j)$ is a 'trend effect' which applies at time t_j, $\phi_{h(i)}$ is a random subject intercept, $\varepsilon_{h(i)j}$ is a residual within-patient 'disturbance' term assumed independent between but not necessarily within patients, $\tau_h(t_j)$ is some 'average' effect of treatment h at time t_j and $\gamma_{h(i)}(t_j)$ is a corresponding patient by treatment interaction term or random coefficient. (This model is overparameterized and interest will centre on identifiable contrasts of $\tau_h(t_j)$.) It is the form of $\tau_h(t_j)$ which determines what sort of summary measure should be used. If we believe that this is well approximated by

$$\tau_h(t_j) = \beta_h t_j \tag{4}$$

then slopes are an appropriate summary measure to use. If, on the other hand, we believe that the appropriate function is

$$\tau_h(t_j) = \beta_h, \quad j \geqslant 1 \tag{5}$$

then means are appropriate. (Once the summary measures have been calculated, we may then proceed to use these to estimate $\beta_2 - \beta_1$ in either case.) Of course, the choice of model is by no

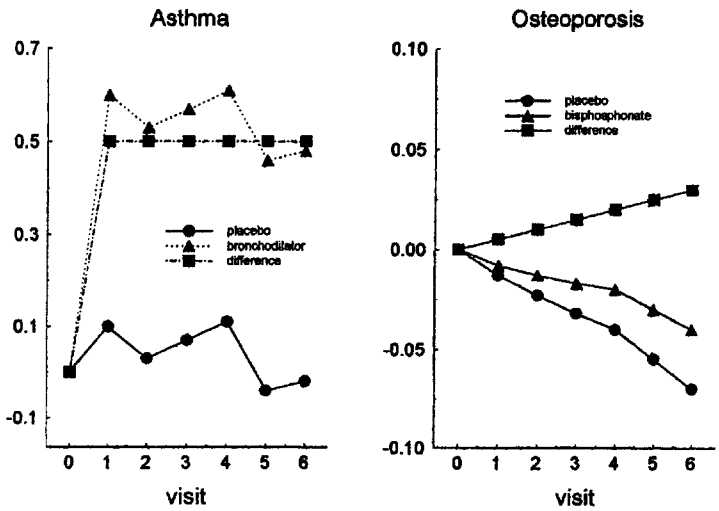


Figure 1. Treatment effects in asthma and osteoporosis. For asthma, difference from baseline in FEV_1 in litres is shown. For osteoporosis, the scale is percentage reduction in bone mineral density.

means restricted to the two forms given by (4) and (5) and other summary measures apart from slopes and means are sometimes employed. For example, in pharmacokinetic studies, area under the concentration time curve (AUC) is often employed and also the maximum concentration ($C_{\max}$). The former is very similar to the mean.

In practice, of course, neither (4) nor (5) may obtain. Inspection of data may help determine an appropriate statistic. There are two points of caution to note in this connection. The first is that it is the form of $\tau_h(t_j)$ over time and *not* the form of $Y_{h(i)j}$ which determines the appropriate choice of summary measure [10]. However, the value of $Y_{h(i)j}$ will also reflect the influence of $\pi(t_j)$ and, for the linear model, this is irrelevant to the purpose of determining the choice of summary measure to be used: neither the variance within groups, nor the expected value of the estimate of the treatment effect will be affected by $\pi(t_j)$. Hence, it is not generally appropriate to determine the summary measure solely by examining either individual profile curves or mean curves per treatment group. Instead, it is best to study an appropriate contrast between groups. In this context, the independent analysis referred to above can often be a useful step in determining a final analysis. (This is illustrated in Figure 1, where in both cases the mean traces per treatment group are subject to some irrelevant time influences eliminated in the curve of the difference.) The second point, of course, is that determining the form of summary measure entirely on an inspection of the data can lead to over-fitting.

In practice, whatever summary measure is chosen will be just that: a summary rather than the whole story. Neither means nor slopes are likely to describe perfectly the way in which the effect of treatment varies with time. In this connection it is useful to distinguish two common purposes of summaries. The first is to describe what the effect of treatment was over the course of a study. The second is to describe what effect the treatment had attained by the end of the study. These two rather different purposes can often be satisfied by the same summary measure. For example, if the treatment effect is described by (5) then the mean of all outcome measures may just as well be used

as an estimate of the effect at the end of the trial as a summary of the effect during the trial; in both cases β is being estimated. Similarly, if the treatment effect grows linearly over time, then the effect at the end of the trial is βt_k and the average effect over the course of the trials is $\beta t_k/2$. Clearly, one is simply twice the other and one might just as well report the slope estimate β. Where, however, one is uncertain as to the exact form of $\tau_h(t_j)$, rather different summaries might be used depending on purpose. For example, one might like to give rather more weight to final observations if the purpose is to describe the effect of treatment by the end of the trial rather than over the whole course.

If one believes that the form given by (4) rather than by (5) is appropriate, then a popular choice of summary is the ordinary least squares (OLS) estimate

$$\hat{\beta}_{h(i)} = \frac{\sum_{j=0}^{k} (t_j - \bar{t}.) Y_{h(i)j}}{\sum_{j=0}^{k} (t_j - \bar{t}.)^2} \tag{6}$$

This can then be used as an outcome measure in the second stage in an analysis of covariance, fitting either baselines or estimated intercepts in an analysis of covariance, as discussed by Laird and Wang [7] and Frison and Pocock [2,11].

A disadvantage of this measure, however, is that post-treatment values can enter with different signs in the same treatment group. This makes it vulnerable to departures from linearity (see *Statistical Issues in Drug Development* [10], chapter 8). Consider, for example, the case in which seven equally spaced measurements (including the baseline) are taken. Then (6) is proportional to $-3Y_{h(i)0} - 2Y_{h(i)1} - Y_{h(i)2} + Y_{h(i)4} + 2Y_{h(i)5} + 3Y_{h(i)6}$. The first two measurements post-treatment enter with negative sign. If we have mistaken the form of the treatment effect and it is nearly immediate and then constant as in (5), then the form of adjustment given is most unfortunate. In fact, such OLS estimates are potentially subject to all the disadvantages of analyses that correct for post-treatment measures.

ARE COMPROMISE MEASURES AVAILABLE?

An alternative slope estimate to (6) is given by the regression through the origin estimate. Since in practice the baseline will be fitted in an analysis of covariance we can ignore the contribution of the baseline to this estimate and define it as

$$\hat{\beta}_{h(i)} = \frac{\sum_{j=1}^{k} t_j Y_{h(i)j}}{\sum_{j=1}^{k} t_i^2} \tag{7}$$

Thus, in the example of the seven equally spaced measurements we would be correcting a measure proportional to $Y_{h(i)1} + 2Y_{h(i)2} + 3Y_{h(i)3} + 4Y_{h(i)4} + 5Y_{h(i)5} + 6Y_{h(i)6}$ by the measure $Y_{h(i)0}$ in an analysis of covariance. The measure defined by (7) has the intriguing and attractive property that it can act as both a mean and a slope estimate. In fact, since we can write (6) as

$$\hat{\beta}_{h(i)} = \frac{\sum_{j=1}^{k} t_j Y_{h(i)j}}{\sum_{j=0}^{k} (t_j - \bar{t})^2} - \frac{\bar{t} \sum_{j=1}^{k} Y_{h(i)j}}{\sum_{j=0}^{k} (t_j - \bar{t})^2} - cY_{h(i)0} \tag{8}$$

where c is some suitable constant, since the second term on the right hand side of (8) is simply proportionate to the mean of the measures and since the third term is irrelevant if we use the

baseline measure in an analysis of covariance, it is obvious by inspection that we can construct (7) as a weighted sum of the OLS slope estimate and the mean of all the measures.

AN EXAMPLE

A multi-centre randomized double-blind placebo-controlled trial was performed to assess whether early-stage intervention of an ACE inhibitor (an anti-hypertensive agent) in insulin dependent diabetic patients, without hypertension, would limit the progression of renal disease. This effect was assessed in 530 randomized patients by measuring the degree of albuminuria (protein in the urine). The main outcome measure was albumin excretion rate (AER), which was measured at five equally spaced time points over a period of 24 months, that is, baseline, 6, 12, 18 and 24 months. Because, over time, albumin excretion rate takes on a growth type curve, with a substantial plateau period, and since the RTO estimate uses all post-randomization measurements, gives increased weights to the outcome measurements as the trial proceeds and is robust to early change, it seemed a more suitable choice of summary measure to use than the OLS estimate or the mean of the post-randomization measurements (MEAN).

Owing to the skewness of the distribution of albumin excretion, the log values of albumin excretion were used to calculate the RTO estimate. In calculating these measures one possibility, as already discussed, is to correct the values that go into their calculation by first subtracting the baseline and working with the change scores. For complete data, this makes no difference to any of the analyses provided that the baseline is fitted in an analysis of covariance. Where data are missing it makes no difference to analyses using OLS or mean measures; in the former case because the measure is completely unaffected and in the latter case because for each patient it differs only from the raw measure by the covariate being fitted. However, for the RTO measure, there will be a slight difference because the corrected measure differs by a variable fraction of the baseline from the uncorrected measure, the fraction depending on the number of missing values. In what follows, we illustrate these calculations using uncorrected data.

For the calculations below, the time $t_{h(i)j}$ at which the jth measurement was taken on patient i of treatment group h, is coded as $t_{h(i)j} = \text{month}_{h(i)j}/6$, where $\text{month}_{h(i)j}$ is the number of months after treatment began at which the measurement was taken. Most commonly five measurement were taken, at times 0 (baseline) and 6, 12, 18 and 24 months after treatment, so that the values of $t_{h(i)0}$ to $t_{h(i)5}$ would be 0, 1, 2, 3 and 4. If logged AER values are denoted by laer_0 to laer_4, and $n_{h(i)}$ is the number of measurements taken for patient i of group h, the RTO, OLS and MEAN estimates for the patient are calculated as follows:

$$\text{RTO}_{h(i)j} = \frac{\sum_{j=1}^{n_{h(i)}} t_{h(i)j}\text{laer}_{h(i)j}}{\sum_{j=1}^{n_{h(i)}} t_{h(i)j}^2}$$

$$\text{OLS}_{h(i)j} = \frac{\sum_{j=1}^{n_{h(i)}} (t_{h(i)j} - \bar{t})\text{laer}_{h(i)j}}{\sum_{j=1}^{n_{h(i)}} (t_{h(i)j} - \bar{t})^2}$$

$$\text{MEAN}_{h(i)j} = \frac{\sum_{j=1}^{n_{h(i)}} \text{laer}_{h(i)j}}{n_{h(i)}}$$

To demonstrate how to calculate these summary measures by hand and using the statistical package SAS®, a random sample of 20 patients was selected from this multi-centre trial (10 on

active therapy and 10 on placebo), the data of which are given in Appendix A. (It should be noted that this illustrative sample is far too small to say anything useful about this treatment.) Patients 020005 and 021103 had their last visit at 21 months instead of 24 months. Therefore, to account for this the appropriate value of $t = 3.5$ has been used instead of $t = 4$.

Examples of the calculation of RTO, OLS, MEAN are given in the following using data from patients 001108 (with complete equally spaced data), 003130 (with data missing at 18 months) and 020005 (with data at 21 months instead of 24 months):

Patient 001108

$$\text{RTO} = \frac{1 \times 2.18816 + 2 \times 2.05111 + 3 \times 3.16116 + 4 \times 1.83104}{1 + 4 + 9 + 16} = \frac{23.09802}{30} = 0.76993$$

$$\text{OLS} = \frac{-2 \times 2.6083 + -1 \times 2.18816 + 1 \times 3.16116 + 2 \times 1.83104}{4 + 1 + 1 + 4} = \frac{-0.58152}{10} = -0.05815$$

$$\text{MEAN} = \frac{2.18816 + 2.05111 + 3.16116 + 1.83104}{4} = \frac{9.23147}{4} = 2.30787$$

Patient 003130

$$\text{RTO} = \frac{1 \times 2.29259 + 2 \times 1.98928 + 4 \times 2.05452}{1 + 4 + 16} = \frac{14.48923}{21} = 0.68996$$

$$\text{OLS} = \frac{-1.75 \times 2.22167 + -0.75 \times 2.29259 + 0.25 \times 1.98928 + 2.25 \times 2.05452}{3.0625 + 0.5625 + 0.0625 + 5.0625}$$

$$= \frac{-0.48737}{8.75} = -0.0557$$

$$\text{MEAN} = \frac{2.29259 + 1.98928 + 2.05452}{3} = \frac{6.33639}{3} = 2.11213$$

Patient 020005

$$\text{RTO} = \frac{1 \times 4.64313 + 2 \times 2.93888 + 3 \times 3.76550 + 3.5 \times 3.33078}{1 + 4 + 9 + 12.25} = \frac{33.47512}{26} \times 0.25 = 1.27524$$

$$\text{OLS} = \frac{-1.9 \times 4.82103 - 0.9 \times 4.64313 + 0.1 \times 2.93888 + 1.1 \times 3.7655 + 1.6 \times 3.33078}{3.61 + 0.81 + 0.01 + 1.21 + 2.56}$$

$$= \frac{-3.57359}{8.2} = -0.4358$$

$$\text{MEAN} = \frac{4.64313 + 2.93888 + 3.7655 + 3.33078}{4} = \frac{14.67829}{4} = 3.669573$$

The SAS program in Appendix B named SUMMARY.SAS calculates the three summary measures for each patient.

Table II. Results of analysing the example using three different summary measures approaches.

Summary measure	Difference between active therapy and placebo	Percentage difference in 24 months	95 per cent CI
RTO	− 0.008	3.0*	− 60 to 41
OLS	0.109	− 54.9†	− 175 to 13
Mean	− 0.045	4.4‡	− 52 to 40

* Calculated as (1 − exp(4 × RTO))100.
† Calculated as (1 − exp(4 × OLS))100.
‡ Calculated as (1 − exp(MEAN))100.

Once individual summary measures were calculated for each patient, analysis of covariance was used to assess whether there was a difference in the mean summary measures between the active therapy and placebo groups, after adjusting for the logged value of the baseline albumin excretion rate ($laer_0$). The difference between the adjusted mean RTO (95 per cent CI) of the active therapy and placebo groups was − 0.008 (− 0.132, 0.117) and can be interpreted as the difference in the rate of change of log(AER) between the active therapy and placebo groups in a 6-month period. The equivalent difference for the OLS was 0.109 (− 0.034, 0.252) and for the MEAN was − 0.045 (− 0.508, 0.418).

As stated previously, these absolute differences in the summary measures relate to the log values AER, which amount to relative differences for the actual albumin excretion rates. Hence, to interpret these data in terms of the actual AER values, the relative change in AER of active therapy to placebo over a 24-month period was calculated. This was calculated for the RTO as (1 − exp(4ΔRTO))100, for OLS as (1 − exp(4ΔOLS))100 and for the MEAN as (1 − exp(ΔMEAN))100, where ΔRTO, ΔOLS, ΔMEAN are the differences in the adjusted mean values of the RTO, OLS and MEAN calculated using analysis of covariance. The results of analysis of covariance using the RTO, OLS and MEAN as summary measures are presented in Table II.

CONCLUDING RECOMMENDATIONS

The summary measures approach is a simple and robust approach to analysing clinical trials. In many cases the loss of efficiency compared to fitting more formal hierarchical models is not great. Just as with other approaches, however, model choices have to be made. Common measures are slopes and means. The choice of measure should depend on the way in which the effect of treatment is believed to change over time but not on general trends. An attractive compromise between mean and slope estimates is the regression through the origin estimate. Baseline measurements should be fitted in analysis of covariance. If more than one baseline measurement is available and if the baseline observation period is short compared to the rest of the trial, then it will usually be appropriate to fit the mean baseline. Where the baseline period of observation is long, it may be advantageous to fit all baselines as covariates. It will usually not be efficient to fit the last baseline only.

APPENDIX A: DATA IN THE SAS DATA SET VISIT.SD2

OBS	PATIENT	T	LAER	LAERO	THERAPY
1.00	1008	0.000	2.608	2.608	Active
2.00	1108	1.000	2.188	2.608	Active
3.00	1108	2.000	2.051	2.608	Active
4.00	1108	3.000	3.161	2.608	Active
5.00	1108	4.000	1.831	2.608	Active
6.00	3130	0.000	2.222	2.222	Active
7.00	3130	1.000	2.293	2.222	Active
8.00	3130	2.000	1.989	2.222	Active
9.00	3130	4.000	2.055	2.222	Active
10.00	7133	0.000	2.629	2.629	Placebo
11.00	7133	1.000	3.288	2.629	Placebo
12.00	7133	2.000	2.575	2.629	Placebo
13.00	7133	3.000	1.985	2.629	Placebo
14.00	7133	4.000	2.297	2.629	Placebo
15.00	8128	0.000	2.437	2.437	Active
16.00	8128	1.000	1.226	2.437	Active
17.00	8128	2.000	1.959	2.437	Active
18.00	8128	3.000	3.313	2.437	Active
19.00	8128	4.000	2.367	2.437	Active
20.00	10017	0.000	3.177	3.177	Active
21.00	10017	1.000	1.684	3.177	Active
22.00	10017	2.000	2.765	3.177	Active
23.00	10017	3.000	1.058	3.177	Active
24.00	10017	4.000	2.723	3.177	Active
25.00	11135	0.000	0.666	0.666	Placebo
26.00	11135	1.000	1.361	0.666	Placebo
27.00	11135	2.000	1.518	0.666	Placebo
28.00	11135	3.000	1.624	0.666	Placebo
29.00	11135	4.000	1.470	0.666	Placebo
30.00	15024	0.000	3.530	3.530	Placebo
31.00	15024	1.000	1.626	3.530	Placebo
32.00	15024	2.000	2.556	3.530	Placebo
33.00	15024	3.000	1.754	3.530	Placebo
34.00	15125	0.000	2.892	2.892	Active
35.00	15125	1.000	1.912	2.892	Active
36.00	15125	3.000	3.517	2.892	Active
37.00	15125	4.000	3.958	2.892	Active
38.00	15136	0.000	2.429	2.429	Placebo
39.00	15136	1.000	3.697	2.429	Placebo
40.00	15136	2.000	2.676	2.429	Placebo
41.00	15136	3.000	2.079	2.429	Placebo
42.00	15136	4.000	2.176	2.429	Placebo
43.00	16024	0.000	2.155	2.155	Placebo
44.00	16024	1.000	1.581	2.155	Placebo
45.00	16024	2.000	1.804	2.155	Placebo
46.00	16024	3.000	1.666	2.155	Placebo
47.00	16024	4.000	2.124	2.155	Placebo
48.00	16101	0.000	1.146	1.146	Active
49.00	16101	1.000	0.507	1.146	Active
50.00	16101	2.000	1.506	1.146	Active
51.00	16101	3.000	1.236	1.146	Active

APPENDIX A: (Continued)

52.00	16101	4.000	1.185	1.146	Active
53.00	16134	0.000	2.330	2.330	Placebo
54.00	16134	1.000	0.449	2.330	Placebo
55.00	16134	2.000	1.593	2.330	Placebo
56.00	16134	3.000	1.261	2.330	Placebo
57.00	16134	4.000	1.666	2.330	Placebo
58.00	17102	0.000	1.052	1.052	Placebo
59.00	17102	1.000	0.733	1.052	Placebo
60.00	17102	2.000	0.741	1.052	Placebo
61.00	17102	3.000	1.367	1.052	Placebo
62.00	17102	4.000	1.588	1.052	Placebo
63.00	17107	0.000	0.803	0.803	Placebo
64.00	17107	1.000	−0.099	0.803	Placebo
65.00	17107	2.000	1.195	0.803	Placebo
66.00	17107	3.000	0.999	0.803	Placebo
67.00	17107	4.000	2.038	0.803	Placebo
68.00	17114	0.000	2.884	2.884	Placebo
69.00	17114	1.000	3.685	2.884	Placebo
70.00	17114	2.000	3.036	2.884	Placebo
71.00	17114	3.000	2.365	2.884	Placebo
72.00	17114	4.000	3.319	2.884	Placebo
73.00	17118	0.000	2.115	2.115	Active
74.00	17118	1.000	2.079	2.115	Active
75.00	17118	2.000	2.834	2.115	Active
76.00	17118	3.000	2.479	2.115	Active
77.00	17118	4.000	1.624	2.115	Active
78.00	20005	0.000	4.821	4.821	Active
79.00	20005	1.000	4.643	4.821	Active
80.00	20005	2.000	2.939	4.821	Active
81.00	20005	3.000	3.766	4.821	Active
82.00	20005	3.500	3.331	4.821	Active
83.00	20101	0.000	1.696	1.696	Active
84.00	20101	1.000	−0.195	1.696	Active
85.00	20101	2.000	1.120	1.696	Active
86.00	20101	3.000	0.370	1.696	Active
87.00	20101	4.000	1.572	1.696	Active
88.00	20116	0.000	1.476	1.476	Placebo
89.00	20116	1.000	2.106	1.476	Placebo
90.00	20116	2.000	1.807	1.476	Placebo
91.00	20116	3.000	1.151	1.476	Placebo
92.00	20116	4.000	1.482	1.476	Placebo
93.00	21103	0.000	2.196	2.196	Active
94.00	21103	1.000	1.762	2.196	Active
95.00	21103	2.000	1.733	2.196	Active
96.00	21103	3.000	2.456	2.196	Active
97.00	21103	3.500	1.642	2.196	Active

Data list

PATIENT	Patient number
T	Time of visit: 0, baseline; 1, 6 months; 2, 12 months, 18 months; 3.5, 21 months; 4, 24 months

LAER	Log of AER
LAERO	Log of AER
LAER0	Log of AER at baseline
THERAPY	Therapy group: active or placebo

APPENDIX B. THE SAS PROGRAM SUMMARY.SAS TO CALCULATE THE SUMMARY MEASURES RTO, OLS AND MEAN FOR INDIVIDUAL PATIENTS USING THE SAS DATA SET VISIT.SD2

```
*Name library;

libname trial 'a:';
run;

*calculate individual rto for all patients;

proc reg data = trial.visit outest = rto noprint; model laer = t/noint;
by patient;
data trial.rto; set rto; keep patient t;
if t ne 0;
if t ne.;
rename t = rto;
proc sort; by patient;
run;

*calculate ols estimate of beta for each patient;

proc reg data = trial.visit outest = ols noprint; model laer = t;
by patient;
data trial.ols; set ols; keep patient t;
if t ne 0;
if t ne.;
rename t = ols;
proc sort; by patient;
run;

*calculate mean ln(aer) for followup visits;

data mean; set trial.visit;
if t ne 0;
proc sort; by patient;
proc means noprint; var laer; by patient;
proc means noprint; var laer; by patient;
output out = trial.mean mean = mean;
run;
```

REFERENCES

1. Matthews JNS, Altman DG, Campbell MJ, Royston P. Analysis of serial measurements in medical research. *British Medical Journal* **300**:230–235.
2. Frison L, Pocock S. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Statistics in Medicine* 1992; **11**:1685–1704. (Correction see Senn SJ. Letter to the editor. *Statistics in Medicine* 1994; **13**:197–198.)
3. Feldman HA. Families of lines: random effects in linear regression analysis. *Journal of Applied Physiology* 1988; **64**:1721–1732.
4. O'Brien P. Procedures for comparing samples with multiple endpoints. *Biometrics* 1984; **40**:1079–1087.
5. Finney D. Repeated measures: what is measured and what repeats. *Statistics in Medicine*, 1990; **9**:639–644.
6. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **35**:795–802.
7. Laird N, Wang F. Estimating rates of change in randomised clinical trials. *Controlled Clinical Trials* 1990; **11**:405–419.
8. Diggle P, Liang KY, Zeger SL. *Analysis of Longitudinal Data.* Oxford Science Publications: Oxford, 1994.
9. Laird N. Further comparative analyses of pre-test post-test research design. *American Statistician* 1983; **37**:329–330.
10. Senn SJ. *Statistical Issues in Drug Development.* Wiley: Chichester, 1997.
11. Frison L, Pocock S. Linearly divergent treatment effects in clinical trials with repeated measures: efficient analysis using summary statistics. *Statistics in Medicine* 1997; **16**:2855–2872.

TUTORIAL IN BIOSTATISTICS

Strategies for comparing treatments on a binary response with multi-centre data

Alan Agresti*,† and Jonathan Hartzel

Department of Statistics, University of Florida, Gainesville, Florida 32611-8545, U.S.A.

SUMMARY

This paper surveys methods for comparing treatments on a binary response when observations occur for several strata. A common application is multi-centre clinical trials, in which the strata refer to a sample of centres or sites of some type. Questions of interest include how one should summarize the difference between the treatments, how one should make inferential comparisons, how one should investigate whether treatment-by-centre interaction exists, how one should describe effects when interaction exists, whether one should treat centres and centre-specific treatment effects as fixed or random, and whether centres that have either 0 successes or 0 failures should contribute to the analysis. This article discusses these matters in the context of various strategies for analysing such data, in particular focusing on special problems presented by sparse data. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

For motivation regarding the questions addressed in this paper, we begin with an example. Table I shows results of a clinical trial conducted at eight centres. The purpose was to compare two cream preparations, an active drug and a control, with respect to their success or failure in curing an infection [1]. This table illustrates a common situation in many pharmaceutical and biomedical applications – comparison of two treatments on a binary response ('success' or 'failure') when observations occur for several strata. The strata are often medical centres or clinics, or they may be levels of a control variable, such as age or severity of the condition being treated, or combinations of levels of several control variables, or they may be different studies of the same sort evaluated in a meta analysis [2–10].

Table I exhibits a potential difficulty that often occurs with multi-centre clinical trials or stratification using several control variables: the sample sizes for the treatments in many of the clinics are modest, and the corresponding cell counts are relatively small. Indeed, for the control group

Correspondence to: Alan Agresti, Department of Statistics, University of Florida, Gainesville, Florida 32611-8545, U.S.A.
†E-mail: AA@STAT.UFL.EDU

Contract/grant sponsor: NIH
Contract/grant sponsor: NSF

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
 ISBN: 0-470-02365-1

Table I. Clinical trial relating treatment to response for eight centres.

Centre	Treatment	Response		Total	Per cent 'success'
		Success	Failure		
1	Drug	11	25	36	30.6
	Control	10	27	37	27.0
2	Drug	16	4	20	80.0
	Control	22	10	32	68.8
3	Drug	14	5	19	73.7
	Control	7	12	19	36.8
4	Drug	2	14	16	12.5
	Control	1	16	17	5.9
5	Drug	6	11	17	35.3
	Control	0	12	12	0.0
6	Drug	1	10	11	9.1
	Control	0	10	10	0.0
7	Drug	1	4	5	20.0
	Control	1	8	9	11.1
8	Drug	4	2	6	66.7
	Control	6	1	7	85.7
Total	Drug	55	75	130	42.3
	Control	47	96	143	32.9

Source of data: Beitler and Landis [1].

in two centres, all observations are failures. Ordinary maximum likelihood (ML) estimation can provide badly biased (even infinite) estimates of some parameters in such cases, and in certain asymptotic frameworks it can even be inconsistent. Bias also occurs, however, from combining strata to increase the stratum-specific sample sizes.

Among the questions of interest for data of this sort are the following: (i) How should one summarize, descriptively, the difference between the treatments? (ii) How should one make inferential comparisons of the treatments? (iii) How should one investigate whether there is treatment-by-centre interaction? (iv) If such interaction exists, how should one describe the effect heterogeneity? (v) Should centres be treated as fixed or random, and could that choice affect any results in a substantive way? (vi) Should centres with 0 successes or with 0 failures contribute to descriptive and inferential analyses? (vii) Should one combine centres or add small constants to empty cells in descriptive and inferential analyses, for instance to use information that otherwise is discarded in the statistical analysis?

In considering these questions, this article discusses strategies for analysing data of the form of Table I. Section 2 presents some possible models for the data and corresponding summaries of the effects. Section 3 presents ways of estimating those effects, and Section 4 illustrates the models for Table I. Section 5 discusses inferential analyses for the models. Section 6 studies the effects of severe sparseness on the analyses, using a data set that is even more sparse than Table I. Section 7 compares the strategies, makes some recommendations, and mentions extensions, alternative approaches, and open questions.

Possible analyses result from all combinations of several factors, including: (i) the choice of link function relating response probabilities to predictors in the model; (ii) whether the model

permits interaction; (iii) whether the model treats centres as random or fixed; (iv) whether inference uses a small-sample analysis or an asymptotic one with the number of centres fixed or an asymptotic one with the number of centres growing with the sample size; (v) whether one uses a Bayes or a frequentist approach or some non-likelihood-based method such as generalized estimating equations (GEE). Here, we consider only the frequentist approach and binary responses. Other papers have presented related discussion of some of these issues in the contexts of Bayesian approaches [4, 6, 11, 12] and continuous responses [13–15]. Also, we do not consider other issues of importance in actual clinical trials, such as adequacy of sample size and selection of centres.

This paper does not claim any new or surprising results, and although it is called a 'tutorial', we fully expect that many readers will have strong opinions about the appropriateness of certain methods. We hope, however, that a unified discussion of various strategies may be helpful for many biostatisticians and quantitatively-oriented medical researchers and perhaps even stimulate research on alternative approaches.

2. MODELS AND SUMMARIES OF EFFECTS

For data in the form of Table I, let X denote treatment, let Y denote the response variable, and let Z denote the stratification factor. Let $X=1$ denote the drug and $X=2$ denote the control (or placebo), and let $Y=1$ denote 'success' and $Y=2$ denote 'failure'. Let $\pi_{ik}=P(Y=1|X=i,Z=k)$, for $i=1,2,k=1,\ldots,K$. Let n_{ijk} denote the cell count for treatment i and response outcome j in stratum k. In this article we often refer to Z using the generic term 'centre', although as mentioned above it might refer to different studies or combinations of levels of control variables.

2.1. Models assuming a lack of interaction

A simple model for Table I, although usually only plausible to a rough approximation, has additive treatment and centre effects on some scale. For instance, with the logit link function (that is, log of the odds) $\text{logit}(\pi_{ik})=\log[\pi_{ik}/(1-\pi_{ik})]$, this is

$$\text{logit}(\pi_{1k})=\alpha_k+\beta/2,\quad \text{logit}(\pi_{2k})=\alpha_k-\beta/2,\ k=1,\ldots,K \tag{1}$$

That is, β is the difference between the logit for drug and the logit for control. One could include an overall intercept in this model and then use a constraint such as $\sum_k \alpha_k=0$ or $\alpha_1=0$, but we use parameterization (1) to discuss more easily (later in the paper) the effects of strata with 0 successes or with 0 failures.

This model assumes a lack of treatment-by-centre interaction. For the logit scale, β refers to a log-odds ratio, so a lack of interaction implies that the true odds ratio e^{β} between X and Y is the same in all centres. Usually primary interest focuses on estimating the treatment effect β rather than the centre effects $\{\alpha_k\}$.

When additivity exists, it need not be on the logit scale. In addition, many practitioners have difficulty interpreting the odds ratio. One could use the same predictor form with an alternative link function such as the probit or log-log or complementary log-log, although these can also be difficult to interpret. Simpler interpretations occur with the log link, by which

$$\log(\pi_{1k})=\alpha_k+\phi/2,\ \log(\pi_{2k})=\alpha_k-\phi/2 \tag{2}$$

With this model, $\exp(\phi) = \pi_{1k}/\pi_{2k}$ is a ratio of success rates, analogous to a relative risk in each centre. (Here, we use notation ϕ rather than β to reflect the effect having a different meaning than in model (1); likewise, the intercept also refers to a different scale, but we use common α_k notation for simplicity since this parameter is not the main focus of interest.)

Model (2) has the structural disadvantage of constraining $\alpha_k \pm \phi/2$ to be negative, so that π_{ik} falls between 0 and 1. Iterative methods for fitting the model may either ignore this, perhaps yielding estimates of some π_{ik} above the permissible $[0, 1]$ range, or may fail to converge if estimates at some stage violate this restriction; normally this does not happen when $\{\pi_{ik}\}$ are not near 1. This model approximates the logit model when $\{\pi_{ik}\}$ are close to 0, but it has interpretations for ratios of probabilities rather than ratios of odds. Model (2) refers to a ratio of success rates, and unlike other models considered in this subsection, when it holds it no longer applies if one interchanges the labelling of 'success' and 'failure' categories.

Simple interpretations also occur with the identity link, by which

$$\pi_{1k} = \alpha_k + \delta/2, \;\; \pi_{2k} = \alpha_k - \delta/2 \tag{3}$$

For this model, the probability of success is $\pi_{1k} - \pi_{2k} = \delta$ higher for drug than control in each centre. This model has the severe constraint that $\alpha_k \pm \delta/2$ must fall in $[0, 1]$. Iterative methods often fail for it. It is unlikely to fit well when any π_{ik} are near 0 or 1 as well as somewhat removed from those boundary values, since smaller values of $\pi_{1k} - \pi_{2k}$ typically occur near the parameter space boundary. Thus, the model has less scope than the ones with logit and log links. Even so, unless the model fits very poorly, an advantage of summarizing the effect by δ is its ease of interpretation by non-statisticians.

In summarizing association for a set of centres by a single measure such as the odds ratio or relative risk, it is preferable to use the measure that is more nearly constant across those centres. In practice, however, for sparse data it is usually difficult to establish superiority of one link function over others, especially when all $\{\pi_{ik}\}$ are close to 0. This article discusses all three of these link functions but pays greatest attention to the logit, which is the most popular one in practice.

2.2. *Random effects models*

The standard ML approach for fitting models such as (1) treats $\{\alpha_k\}$ as fixed effects. In many applications, such as multi-centre clinical trials and meta analyses, the strata are themselves a sample. When this is true and one would like inferences to apply more generally than to the strata sampled, a random effects approach may be more natural. In practice, the sample of strata are rarely randomly selected. However, Grizzle [16] expressed the belief of many statisticians when he argued that 'Although the clinics are not randomly chosen, the assumption of random clinic effect will result in tests and confidence intervals that better capture the variability inherent in the system more realistically than when clinic effects are considered fixed'. This approach seems reasonable to us for many applications of this type.

For the logit link, a logit-normal random effects model [17] with the same form as (1)

$$\text{logit}(\pi_{1k}) = a_k + \beta/2, \quad \text{logit}(\pi_{2k}) = a_k - \beta/2 \tag{4}$$

assumes that $\{a_k\}$ are independent from a $N(\alpha, \sigma)$ distribution. The parameter σ, itself unknown, summarizes centre heterogeneity in the success probabilities. This model also makes the strong assumption that the treatment effect β is constant over strata.

For binary data, random effects models are most commonly used with logit or probit link functions. A structural defect exists with the log and identity links in treating $\{a_k\}$ as normally distributed; for any parameter values with $\sigma > 0$, with positive probability a particular realization of the random effect corresponds to π_{ik} outside $[0, 1]$.

2.3. *Treatment-by-centre interaction*

Even if a model that is additive in centre and treatment effects fits sample data adequately, it is usually unrealistic to expect the *true* association to be identical (or essentially identical) in each stratum. This subsection considers models that permit interaction. With a fixed-effects approach, the model

$$\text{logit}(\pi_{1k}) = \alpha_k + \beta_k/2, \quad \text{logit}(\pi_{2k}) = \alpha_k - \beta_k/2 \tag{5}$$

has odds ratio e^{β_k} in centre k. It is saturated (residual d.f. $= 0$), having $2K$ parameters for the $2K$ binomial probabilities. The ML estimate of β_k is the sample log-odds ratio in stratum k, $\hat{\beta}_k = \log(n_{11k}n_{22k}/n_{12k}n_{21k})$.

Usually, such as in meta analyses, one would want to extend such a model to determine explanations for the variability in associations among the strata. When the strata have a natural ordering with scores $\{z_k\}$, an unsaturated model (d.f. $= K - 2$) results from assuming a linear trend in the log-odds ratios; that is, by replacing β_k in model (5) by $\beta + z_k\lambda$. Often other explanatory variables are available for modelling the odds ratio [18–20]. Then, one could construct a model of form

$$\beta_k = \mathbf{z}_k'\boldsymbol{\lambda}$$

describing the centre-specific log-odds ratios, where $\mathbf{z}_k$ is a column vector of explanatory variables and $\boldsymbol{\lambda}$ is a column vector of parameters. A related model adds a random effect term for each centre to reflect unexplained variability [21].

For the random effects approach without other explanatory variables, an additional parameter can represent variability in the true effects. The logit-normal model is

$$\text{logit}(\pi_{1k}) = a_k + b_k/2, \quad \text{logit}(\pi_{2k}) = a_k - b_k/2 \tag{6}$$

where $\{a_k\}$ are independent from $N(\alpha, \sigma_a)$, $\{b_k\}$ are independent from $N(\beta, \sigma_b)$, and $\{a_k\}$ are independent of $\{b_k\}$. Here, β is the expected value of centre-specific log-odds ratios, and σ_b describes their variability. An equivalent model form is $\text{logit}(\pi_{ik}) = a_k + \beta x_i + b_{ik}$, where x_i is a treatment dummy variable ($x_1 = 1, x_2 = 0$) and b_{1k} and b_{2k} are independent $N(0, \sigma)$, where σ^2 corresponds to $\sigma_b^2/2$ in parameterization (6). Note that one should not formulate the model as $\text{logit}(\pi_{ik}) = a_k + b_k x_i$, since the model then imposes greater variability on the logit for the first treatment unless one permits (a_k, b_k) to be correlated.

Analogous random effects models apply with alternative link functions. Again, the models with identity or log link are structurally improper when either variance component is positive. This suggests a caution, as results reported for software using a particular estimation method may depend on whether the parameter constraints are recognized. In our experience, the identity link often has convergence problems. Good initial estimates of $(\alpha, \beta, \sigma_a, \sigma_b)$ can be helpful, such as using values suggested by fixed effects modelling. In some applications it is also sensible to let (a_k, b_k) be correlated, by treating it as a bivariate normal random effect [22]. With the identity link, for instance, centres with a_k close to 0 may tend to have values of b_k relatively close to

0. We do not discuss such models here, as such modelling is better supported with moderate to large K, and our examples have relatively small K with sparse data. With such examples, some will think it bold or foolhardy of us to use even relatively simple random effects models!

3. MODEL FITTING AND ESTIMATING EFFECTS

We now discuss model fitting and parameter estimation. Unless stated otherwise, the discussion refers to the logit models.

3.1. *Model fitting*

It is straightforward to fit the fixed effects models with standard software. Possibilities include software for binary responses such as PROC LOGISTIC in SAS, or software for generalized linear models such as PROC GENMOD in SAS and the *glm* function in S-plus.

Random effects models for binary data are more difficult to fit. One must integrate the joint mass function of the responses with respect to the random effects distributions to obtain the likelihood function [23], which is a function of β and the other parameters of those distributions. With the logit interaction model (6), for instance, the likelihood function equals

$$\ell(\alpha, \beta, \sigma_a, \sigma_b) = \Pi_k \Pi_i \left[\int_{a_k} \int_{b_k} \pi_{ik}^{n_{i1k}} (1 - \pi_{ik})^{n_{i2k}} \, \mathrm{d}G(b_k) \, \mathrm{d}F(a_k) \right]$$

where F is a $\mathrm{N}(\alpha, \sigma_a)$ CDF, G is a $\mathrm{N}(\beta, \sigma_b)$ CDF, $\pi_{1k} = \exp(a_k + b_k/2)/[1 + \exp(a_k + b_k/2)]$, and $\pi_{2k} = \exp(a_k - b_k/2)/[1 + \exp(a_k - b_k/2)]$. One can approximate the likelihood function using numerical integration methods, such as Gauss–Hermite quadrature. The approximation improves as the number of quadrature points q increases, more points being needed as the variance components increase in size. Performance is enhanced by an adaptive version of quadrature that transforms the variable of integration so that the integrand is sampled in an appropriate region [24, 25]. Having approximated the likelihood, one can use standard maximization methods such as Newton–Raphson to obtain the estimates. As a by-product, the observed information matrix, based on the curvature (second derivatives) of the log-likelihood at the ML estimates, is inverted to provide an estimated asymptotic covariance matrix.

Other approximations for integrating out the random effects lead to related approximations of the likelihood function and the ML estimates. Most of these utilize linearizations of the model. A Laplace approximation yields penalized quasi-likelihood (PQL) estimates [26], and a related generalization includes an extra scale parameter [27]. These approximations can behave poorly when variance components are large or when distributions are far from normal, such as Bernoulli or binomial with small indices at each setting of predictors [25, 26, 28]. When feasible, it is better to use adaptive Gauss–Hermite quadrature with sufficiently large q, the determination of 'sufficiently large' being based on monitoring the convergence of estimates and standard errors as q increases. Other promising ML approximations use Monte Carlo approximation methods [28, 29], for which the approximation error is estimable and decreases as the number of simulations increases. One can also use Markov chain Monte Carlo methods with an approximating Bayes model that uses flat prior distributions for the other parameters [30], although the danger exists of improper posterior distributions [31–33].

Most major software packages are not equipped to fit generalized linear models with random effects. Version 7 of SAS includes PROC NLMIXED, which can provide a good approximation to ML using adaptive Gauss–Hermite quadrature. The linearization approximations [26, 27] are available in earlier versions with a SAS macro, called GLIMMIX, that uses iterative calling of PROC MIXED. Most other specialized programs for hierarchical models with random effects likewise use various normal approximations to the working response in the mixed logit model.

3.2. The sparse asymptotic framework

In many applications, such as when the strata are centres, asymptotic arguments for increasing the sample size most naturally refer to increasing simultaneously the number of strata, K. A disadvantage then of the usual large-sample methods with the fixed effects logit models is that they are based on $n \to \infty$ with a *fixed* number of parameters (for example, K fixed), whereas the more appropriate 'sparse asymptotic' framework has $K \to \infty$ as $n \to \infty$. For sparse asymptotics, consistency of ordinary ML estimators breaks down for the odds ratio, relative risk, and difference of proportions [34]. An extreme case (Anderson [35], p. 244) occurs with matched-pairs data (two observations for each k), in which case the ordinary ML estimator of β in model (1) converges in probability to 2β.

The sparse asymptotic framework does not cause special problems for the random effects approach. After integrating out the random effects, the likelihood function depends only on the remaining parameters (for example, α, β and σ in model (4)), so the parameter space does not increase as K does. In particular, if the random effects model holds, the ordinary ML estimator of β is consistent. In practice, however, if n and K have only moderate size, as in Table I, inferences about the size of the variance components may be very imprecise.

In the logit fixed effects model (1), the conditional likelihood approach provides an alternative way of guaranteeing a consistent estimator of β. With it, one eliminates $\{\alpha_k\}$ in constructing the likelihood function by conditioning on their sufficient statistics [36]. Software is available for this approach, such as LogXact [37]. It has the advantage of not requiring a distributional assumption about the random effects yet still being valid for sparse asymptotics. A disadvantage of conditional ML is that the fitting procedure does not provide predicted values for $\{\alpha_k\}$ or an estimate of their variability. Also, this approach is applicable only with the logit link (that is, only the canonical link of a generalized linear model provides reduced sufficient statistics).

3.3. Mantel–Haenszel type estimators of common effects

An alternative estimator of β in the no interaction model (1) is the Mantel–Haenszel (M–H) estimator [38]

$$\hat{\beta}_{\mathrm{MH}} = \log\left(\frac{\sum_k n_{11k} n_{22k}/n_{++k}}{\sum_k n_{12k} n_{21k}/n_{++k}}\right) \tag{7}$$

Like the conditional ML estimator, it is consistent both in sparse-stratum (K increases with n) or large-stratum (K fixed but n increases) asymptotics. It has the advantage over conditional ML of simplicity. It suffers no efficiency loss when $\beta = 0$ and usually little otherwise.

Mantel–Haenszel type estimators are also available for the relative risk and the difference of proportions. As noted in Section 2, models for these parameters have severe parameter restrictions.

Even if the model holds only approximately, however, a summary measure of this type is useful for communication with scientists who are unfamiliar with odds ratios. The M–H type estimator of a common log relative risk [39, 40] (that is, ϕ in model (2)) is

$$\hat{\phi}_{\mathrm{MH}} = \log\left(\frac{\sum_k n_{11k}n_{2+k}/n_{++k}}{\sum_k n_{21k}n_{1+k}/n_{++k}}\right) \tag{8}$$

whereas the M–H type estimator of a common difference of proportions [34] (that is, δ in model (3)) is

$$\hat{\delta}_{\mathrm{MH}} = \frac{\sum_k (n_{11k}n_{2+k}/n_{++k} - n_{21k}n_{1+k}/n_{++k})}{\sum_k n_{1+k}n_{2+k}/n_{++k}} \tag{9}$$

If a no interaction model fits adequately but the data are highly sparse, the corresponding M–H estimator may even be preferred to the ML estimator, because of the bias that exists in sparse asymptotics [34] for the ML estimator. The conditional ML approach does not apply to the log and identity link functions and the random effects model has structural problems (for example, probabilities outside the $[0, 1]$ interval), so these estimates are particularly useful for these link functions.

Given their good performance under sparse asymptotics and their ease of computation, one might consider always using M–H instead of ML estimators. However, for large-stratum asymptotics (fixed K), M–H estimators lose some efficiency compared to ML, and the efficiency loss can be considerable for $\hat{\phi}_{\mathrm{MH}}$ and $\hat{\delta}_{\mathrm{MH}}$ in some cases [34]. Moreover, software for ML estimation is widely available for fixed effects analyses and becoming more so for random effects analyses. Thus, if the data have moderate to large samples in each stratum, it is better to use the model-based ML estimators.

3.4. Centre estimates

In most applications, main interest focuses on the treatment effect and its variability among centres. However, centre estimates also result from the fixed effects or random effects ML approaches. With the random effects approach, the expected values of $\{a_k\}$ given the data are analogues of best linear unbiased predictors (BLUP) for mixed models with normal responses. These expected values themselves depend on unknown parameters, so one obtains the predicted values by plugging in the ML estimates of those parameters. Ordinary standard errors of these predictors, like those of empirical Bayes estimators, do not take into account that the variance component is estimated rather than known; hence, they tend to be too small, and adjustments are available [41, 42]. Adjustments are also available to help account for the bias in estimating the variance components [43], which can be considerable, but we shall not address that issue here.

For fixed effects logit models, the sufficient statistic for α_k is n_{+1k}, conditional on the binomial sample sizes in that stratum. By contrast, for the random effects models estimates of centre effects 'borrow from the whole', and the estimate of a_k can be considerably affected by results in other strata. As the sample size grows in stratum k, however, the influence of other strata decreases.

3.5. Logit model: Allowing interaction

For the random effects model (6) that permits interaction, the complexity of model fitting is compounded by estimating two variance components. When the data are sparse but do contain sufficient

information to provide estimates of $(\alpha, \beta, \sigma_a, \sigma_b)$, the estimated average effects $(\hat{\alpha}, \hat{\beta})$ are more reliable than the estimated variability $(\hat{\sigma}_a, \hat{\sigma}_b)$ of effects, especially when K is not especially large. When $\hat{\sigma}_b > 0$, the standard error of $\hat{\beta}$ is typically larger than with the model (4) of homogeneous odds ratios (that is, the special case in which $\sigma_b = 0$), because of the extra variance component due to treating the treatment effect as random rather than fixed.

Liu and Pierce [44] proposed an alternative way of estimating (β, σ_b) for the model (6) that assumes the log-odds ratios are a $N(\beta, \sigma_b)$ random sample. They first eliminated $\{a_k\}$ by a conditioning argument, focusing solely on the variability in association, and then provided a simple solution based on an approximation to the likelihood function using Laplace's method. They suggested that their method is primarily intended for cases in which cell counts are relatively large and the variability σ_b is not great, say, $\sigma_b < 1$. See Raghunathan and Ii [45] and Liang and Self [46] for related work.

4. MODEL FITTING FOR TABLE I

We now apply these methods to Table I. For these data the sample success rates vary markedly among centres both for the control and drug treatments, but in all except the last centre that rate is higher for drug. Normally in using models with random centre and possibly random treatment effects, one would prefer to have more than $K = 8$ centres; keeping in mind the difficulty particularly of getting good variance component estimates with such a small value of K, we use these data to illustrate the models. Table II shows the use of SAS (PROC NLMIXED and PROC GENMOD) for ML fitting of logit models to Table I. Alternative link functions utilize similar statements. For the random effects interaction model, for instance, the code `pi=exp(a+b*treat)` requests the log link model and `pi=a+b*treat` requests the identity link model. In the NLMIXED code in Table II for the no interaction model with random centre effects, the 'predict' option requests the logit estimates of $a_k \pm \beta/2$ for the eight centres and stores them in the data set OUT1.

Table III summarizes results of estimating the treatment effect β using various logit models. The parameter β is the common log-odds ratio for the no interaction models and the expected value of the log-odds ratio for the interaction model with random treatment effects. For the random effects model (6) permitting interaction, the estimated standard deviation of the log-odds ratios is relatively small, $\hat{\sigma}_b = 0.15$ (standard error $= 1.1$). For all approaches, estimates of the common log-odds ratio or its expected value are similar. In each case the estimated value of about 0.75 equals about 2.5 standard errors; this corresponds to an estimated common odds ratio of about $e^{0.75} = 2.1$ and a 95 per cent confidence interval for the common odds ratio of about (1.2, 3.8). There is considerable evidence of a drug effect, but with such a small sample one cannot determine whether that effect is weak or moderate.

For the interaction model, since $\hat{\sigma}_b$ is small, the random effects model provides a considerable smoothing of the sample odds ratios. Table IV shows the eight sample odds ratios and their random effects model estimates, computed by exponentiating the estimated expected log-odds ratios given the sample data. The smoothed estimates show considerably less variability and do not have the same ordering as the sample values. For instance, the smoothed estimate is greater for centre 3 than for centre 6 even though the sample value is infinite for the latter, partly reflecting the greater shrinkage that occurs when sample sizes are smaller. When $\hat{\sigma}_b = 0$, the interaction model provides the same fit as the no interaction model, so the model estimated odds ratios are identical in each centre.

Table II. Example of SAS code for using GENMOD to fit fixed effects logit model and NLMIXED to fit random effects logit models to Table I.

```
data binomial;
input center treat y n @@ ; * y successes out of n trials;
if treat = 1 then treat = .5; else treat = -.5;
cards;
1 1 11 36    1 0 10 37    2 1 16 20    2 0 22 32
3 1 14 19    3 0  7 19    4 1  2 16    4 0  1 17
5 1  6 17    5 0  0 12    6 1  1 11    6 0  0 10
7 1  1  5    7 0  1  9    8 1  4  6    8 0  6  7
;
run;

proc genmod data = binomial; * fixed effects, no interaction model;
  class center;
  model y/n = treat center / dist = bin link = logit noint;
run;

proc nlmixed data = binomial qpoints = 15; * random effects, no interaction;
  parms alpha = -1 beta = 1 sig = 1; * initial values for parameter estimates;
  pi = exp(a + beta*treat)/(1+exp(a + beta*treat)); * logistic formula for prob;
  model y ~ binomial(n, pi);
  random a ~ normal(alpha, sig*sig) subject = center;
  predict a + beta*treat out = OUT1;
run;

proc nlmixed data = binomial qpoints = 15; * random effects, interaction;
  parms alpha = -1 beta = 1 sig_a = 1 sig_b = 1; * initial values;
  pi = exp(a + b*treat)/(1+exp(a + b*treat));
  model y ~ binomial(n, pi);
  random a b ~ normal([alpha,beta], [sig_a*sig_a,0,sig_b*sig_b]) subject = center;
run;
```

Table III also summarizes estimates for other descriptive measures, with ML results obtained using GENMOD and NLMIXED in SAS. As noted before, the restricted parameter space for the log and identity links can provide problems. Having good starting values increases the chance of proper convergence. We used starting values near the estimates obtained with the SAS GLIMMIX macro.

For the random effects interaction model with the log link, the ML estimated standard deviation of the log relative risks equals 0. Hence, the fitted relative risks are the same in each centre, the estimate of 1.27 being identical to that for the random effects no interaction model. For this sample the association is more nearly constant for the relative risk than the odds ratio.

For the random effects interaction model with the identity link, we were unable to obtain convergence with NLMIXED. Using GLIMMIX, Littell *et al.* [47] reported an estimated mean of 0.120 (standard error = 0.051) and an estimated standard deviation of 0.098 for the clinic-specific differences of proportions, but we could not obtain these results even with GLIMMIX. A weighted least squares estimate of the clinic-specific difference of proportions [8] is 0.131 (standard error = 0.052) with an estimated standard deviation of 0.075.

Table III. Estimated treatment effect and standard error, and results of likelihood ratio (LR) and Wald tests of hypothesis of no treatment effect, for Table I.

Measure (Equation number)	Interaction	Centre	Method	Estimate	Standard error	Wald statistic	LR statistic	*P*-value
Odds ratio (1)	No	Fixed	ML	0.777 (2.2)	0.307	6.4	6.7	0.01
			Cond. ML	0.756 (2.1)	0.303	6.2		
			M–H	0.758 (2.1)	0.304	6.2		
(4)		Random	ML	0.739 (2.1)	0.300	6.0	6.3	0.01
(6)	Yes	Random	ML	0.746 (2.1)	0.325	5.3	4.6	0.03
Relative risk (2)	No	Fixed	ML	0.247 (1.3)	0.126	3.8	3.9	0.05
			M–H	0.354 (1.4)	0.142	6.2		
		Random	ML	0.241 (1.3)	0.126	3.6	3.8	0.05
	Yes	Random	ML	0.241 (1.3)	0.126	3.6	3.7	0.08
Difference of prop. (3)	No	Fixed	ML	0.137	0.055	6.2	6.6	0.01
			M–H	0.130	0.050	6.7		
		Random	ML	0.148	0.055	7.2	7.6	0.01

Odds ratio and relative risk estimates appear in parentheses next to their log estimates. Wald and LR test statistics have approximate null chi-squared distributions with d.f. = 1; *P*-value refers to LR statistic.

Table IV. Estimated centre-specific odds ratio and relative risk for Table I, based on sample and on predictions for random effects interaction models.

Centre	Odds ratio		Relative risk	
	Sample	Model	Sample	Model
1	1.19	2.02	1.13	1.27
2	1.82	2.09	1.16	1.27
3	4.80	2.19	2.00	1.27
4	2.29	2.11	2.13	1.27
5	∞	2.18	∞	1.27
6	∞	2.12	∞	1.27
7	2.00	2.11	1.80	1.27
8	0.33	2.06	0.78	1.27

5. INFERENCE ABOUT EFFECTS

5.1. Inference for logit models

For the fixed and random effects logit models, standard methods yield inferences about the treatment effect. For instance, the likelihood-ratio test statistic is minus twice the difference in maximized log-likelihoods between model (1) or (4) with $\beta = 0$ and the model with unrestricted β. It has a null chi-squared distribution with d.f. = 1, as does the Wald statistic, which is the squared ratio of the estimate to its standard error. The standard error is obtained from the inverse information matrix. The simple Wald form of 95 per cent confidence interval for the common odds ratio is obtained by exponentiating the endpoints of $\hat{\beta} \pm 1.96$(standard error). Better, one could construct a profile likelihood confidence interval (for example, for the fixed effects solution using the LRCI option in PROC GENMOD) or an interval based on inverting a score test [48].

With the fixed effects approach, for highly sparse data it is preferable to conduct inference using the conditional likelihood. For model (1), this likelihood depends only on β. A 95 per cent large-sample likelihood-based confidence interval consists of all β values for which minus two times the log-likelihood falls within 3.84 (the 95th percentile of the χ_1^2 distribution) of the maximum. Tests and confidence intervals with this approach are available with LogXact.

5.2. *Mantel–Haenszel inference*

For model (1), the Mantel–Haenszel estimator (7) of a common log-odds ratio has a standard error estimate [49] that is valid for both large-stratum and sparse-stratum asymptotics. The variance estimate equals

$$\widehat{\text{var}}(\hat{\beta}_{\text{MH}}) = \frac{\sum_k (n_{11k}+n_{22k})(n_{11k}n_{22k})/n_{++k}^2}{2(\sum_k n_{11k}n_{22k}/n_{++k})^2} + \frac{\sum_k (n_{12k}+n_{21k})(n_{12k}n_{21k})/n_{++k}^2}{2(\sum_k n_{12k}n_{21k}/n_{++k})^2} + \frac{\sum_k [(n_{11k}+n_{22k})(n_{12k}n_{21k}) + (n_{12k}+n_{21k})(n_{11k}n_{22k})]/n_{++k}^2}{2(\sum_k n_{11k}n_{22k}/n_{++k})(\sum_k n_{12k}n_{21k}/n_{++k})}$$

One can use this to form a confidence interval for the common log-odds ratio, exponentiating endpoints to obtain the interval for the odds ratio. Like the conditional ML approach, this is preferred over ordinary intervals for the fixed-effects logit model (1) when the data are highly sparse.

Similarly, estimated variances for both types of asymptotics are available for the estimator of a common difference of proportions (9) and the estimator of a common log relative risk (8). Let $R_k = n_{11k}n_{2+k}/n_{++k}$ and $S_k = n_{21k}n_{1+k}/n_{++k}$. For the log relative risk, the estimated variance is [34]

$$\widehat{\text{var}}(\hat{\phi}_{\text{MH}}) = \frac{\sum_k (n_{1+k}n_{2+k}n_{+1k} - n_{11k}n_{21k}n_{++k})/n_{++k}^2)}{(\sum_k R_k)(\sum_k S_k)} \tag{10}$$

For the difference in proportions, the estimated variance is [50]

$$\widehat{\text{var}}(\hat{\delta}_{\text{MH}}) = \frac{\hat{\delta}_{\text{MH}}(\sum_k P_k) + (\sum_k Q_k)}{(\sum_k n_{1+k}n_{2+k}/n_{++k})^2} \tag{11}$$

where

$$P_k = [n_{1+k}^2 n_{21k} - n_{2+k}^2 n_{11k} + n_{1+k}n_{2+k}(n_{2+k} - n_{1+k})/2]/n_{++k}^2$$

and

$$Q_k = [n_{11k}n_{22k} + n_{21k}n_{12k}]/2n_{++k}$$

A disadvantage of these inferences is their restriction to the no interaction models and their treatment effects. Since the variance formulae assume a common treatment effect for each centre, they should not be used when substantial heterogeneity exists.

5.3. *Tests of no interaction*

A test of no interaction for the fixed effects logit model is equivalently a goodness-of-fit test of model (1) and a test for equality of the K true odds ratios. When the data are not sparse and K is fixed, one can use ordinary likelihood-ratio and Pearson chi-squared statistics for this purpose, with

d.f. $= (K-1)$. The likelihood-ratio statistic refers to the likelihood-ratio test comparing the model (1) to the saturated model. An alternative chi-squared test provided by some software for this case is the Breslow–Day test [36, 51], which is based on heterogeneity in the K sample log-odds ratios. For this situation with K fixed, an exact conditional test [52] of equality of odds ratios is available in StatXact [53].

When the stratum-specific sample sizes are small and K is large, none of these tests has much power. It may be possible to increase power by checking for a particular type of interaction, such as a linear trend in the log-odds ratios when the strata have a natural ordering. For the fixed effects approach, a benefit of using the simpler model when the degree of interaction is not significant is that the common odds ratio estimator can be a better estimator of the true stratum-specific odds ratios than the separate sample values (for example, having smaller total mean squared error) even when those true odds ratios are not identical, for the usual reasons of model parsimony.

For the random effects approach, one can test for a lack of interaction by testing that $\sigma_b = 0$ in model (6). The score test with an arbitrary mixture distribution for the random effect leads to an asymptotically normal statistic [46]. Under the null, the likelihood-ratio statistic equals 0 (that is, because $\hat{\sigma}_b = 0$) or approximately a χ^2_1 variate, each with probability about 0.5; thus, the usual chi-squared right-tail probability is halved to get the P-value. However, for random effects models, one might question the entire enterprise of conducting tests of no interaction. Typically the likelihood reveals that values of $\sigma_b > 0$ are consistent with the data, and when $\hat{\sigma}_b > 0$ the confidence interval for β with the interaction model is somewhat wider than with the no interaction model, better reflecting the actual heterogeneity that ordinarily occurs in practice.

5.4. Summarizing effects when interaction exists

When significant interaction exists, with the fixed effects approach the saturated model provides an odds ratio estimate for each stratum. Alternatively, a covariate may be apparent such that odds ratios are more nearly constant after adjusting for that covariate. For instance, there may be one or two centres that are considerably different from the others in some way. With the random effects approach (6), it is natural to describe the interaction by $(\hat{\beta}, \hat{\sigma}_b)$, providing an estimate of an average log-odds ratio and the variability about that average. With the random effects model, one can also obtain approximate BLUP estimates of the log-odds ratios $\{b_k\}$. These provide a smoothing of the sample log-odds ratio estimates from the fixed-effects saturated model. As the sample size increases within a particular stratum, the random effects estimate becomes more similar to the sample value for that stratum.

Similar remarks regarding interaction apply for analyses involving the difference of proportions and relative risk. For example, for fixed K, large-sample d.f. $= K - 1$ chi-squared tests exist of whether the difference of proportions is the same for all strata [8, 48]. A corresponding test holds for sparse data with K large [3]. When interaction exists with a fixed effects model with parameter θ_k in stratum k, an alternative [54–56] to simply reporting the stratum-specific estimates is to estimate $\sum_k \rho_k \theta_k$, where ρ_k is the population proportion classified in stratum k (or if this is unknown, simply $\rho_k = 1/K$).

For the random effects interaction model with identity link, alternative estimates exist of the mean and variance of the stratum-specific differences of proportions [1, 8]. These approaches weight the sample estimate from each stratum inversely proportional to its estimated variance. A modified approach uses an alternative weighting scheme to reduce bias [57]. An analogous random effects analysis exists for the relative risk [9].

5.5. *Goodness-of-fit*

We mentioned above the goodness-of-fit test for the fixed effects models. These treat K as fixed. For sparse data with large K, these tests lack power and may be poorly approximated by chi-squared distributions. Model checking is very challenging with highly sparse data.

The random effects model (4) assuming no interaction also satisfies the fixed effects structure (1). So, lack of fit in the ordinary goodness-of-fit test for model (1) also implies lack of fit in the random effects model. When the random effects model holds, its fit behaves asymptotically like that of the fixed effects model, for fixed K with $n \to \infty$. Similarly, for the models permitting interaction with fixed K, the fixed and random effects estimates are asymptotically equivalent. It is not obvious how to check the fit of such models for sparse asymptotics in which $K \to \infty$. The usual goodness-of-fit statistics are then approximately normal [58], and there is some evidence that the jack-knife can work well in estimating asymptotic variances of such statistics [59]. We are unaware, however, of any checks on this yet for models of the type discussed in this article.

5.6. *Inferential results for Table I*

Table III also shows standard errors for the various estimators and the results of Wald and likelihood-ratio tests of no effect. Substantive results are similar with all link functions, with evidence of a better success rate with drug than with control, although the model-based inferences with the relative risk provide slightly less evidence of association. The estimated effect can be described by a stratum-specific odds ratio of about 2.1, relative risk of about 1.3, or difference of proportions of about 0.14. For each measure the data do not contradict the models that assume a lack of interaction; for instance, the interaction models provide similar summary estimates and standard errors. Also, the traditional goodness-of-fit statistics do not show lack of fit when applied to the fixed effects versions of the no interaction models. The Pearson statistic equals 8.0 for the logit link, 9.9 for the log link, and 9.9 for the identity link, each with d.f. $= 7$.

6. EFFECTS OF SEVERE SPARSENESS

This section summarizes some special considerations and results when the data are severely sparse, such as effects of centres containing certain patterns of empty cells and effects of modifying the data such as by adding constants to empty cells or combining centres. Table V is an example of such data [60]. This table was shown to the first author a few years back by an attendee of a short course on categorical data analysis. It shows results for five centres of a clinical trial designed to compare an active drug to placebo in treating toenail fungal infections. Again, success rates vary markedly among centres, but note that the binomial sample sizes are very small. Here, two centres have no successes and one centre has only one success. Although one cannot expect to conduct precise inference with such small n and K and although normally K would be much larger than 5 in the application of random effects models (especially to estimate variance components), these data are useful for illustrating effects of such severe sparseness.

Here, a reasonable asymptotic framework is the sparse one whereby K increases proportionally to n. When n is small, it is difficult to detect when heterogeneity truly exists among strata in the treatment effects. Thus, our remarks are directed primarily toward models such as (1) and (4), that is, we assume that reality is reasonably well described by the fixed effects or random effects model with homogeneous odds ratios.

Table V. Clinical trial relating treatment to response for five centres.

Centre	Treatment	Response		Total	Per cent 'success'
		Success	Failure		
1	Active drug	0	5	5	0.0
	Placebo	0	9	9	0.0
2	Active drug	1	12	13	7.7
	Placebo	0	10	10	0.0
3	Active drug	0	7	7	0.0
	Placebo	0	5	5	0.0
4	Active drug	6	3	9	66.7
	Placebo	2	6	8	25.0
5	Active drug	5	9	14	35.7
	Placebo	2	12	14	14.3
Total	Active drug	12	36	48	25.0
	Placebo	4	42	46	8.7

Source: Agresti [60], p. 193.

For severely sparse data, the strata sample sizes are very small and using ordinary ML with the fixed effects model may provide seriously biased estimates. If that approach is used, it is safest to do so using conditional ML estimation.

6.1. *Extreme cases: centres with 0 successes or 0 failures*

For stratum k, let $s_k = n_{11k} + n_{21k}$ denote the number of successes and let $f_k = n_{12k} + n_{22k}$ denote the number of failures. First, we study the effects on the analyses of strata that have either $s_k = 0$ or $f_k = 0$, such as centres 1 and 3 of Table V.

Consider fixed effects models relating to the odds ratio. Then, ML estimates exist only in the extended sense that $\hat{\alpha}_k = -\infty$ when $s_k = 0$ and $\hat{\alpha}_k = \infty$ when $f_k = 0$. The likelihood approaches its maximum in the limit as these estimates grow unboundedly in the appropriate direction and $\hat{\beta}$ and $\{\hat{\alpha}_k\}$ for strata with $\min(s_k, f_k) > 0$ take the finite values the ML estimates assume after deleting the offending strata from the data set. Although $\hat{\alpha}_k$ is infinite when $\min(s_k, f_k) = 0$, in practice it is common for software to be fooled by the very flat log-likelihood and converge, reporting large centre estimates. The reported standard errors for such strata are huge, since they are based on inverting a matrix that summarizes the curvature of the log-likelihood at convergence.

In any case, for logit model (1), centres with $s_k = 0$ or $f_k = 0$ have no effect on $\hat{\beta}$. Similarly, the conditional likelihood approach to fitting model (1) ignores strata with $s_k = 0$ or $f_k = 0$, as does the M–H estimate and its standard error. When one conditions on row and column totals, the observed counts in the stratum are the only ones possible, and the distribution is degenerate; for instance, conditionally, the count in the first cell equals the observed value with probability 1, and the variance of the distribution of that count is 0. Similarly, the M–H test statistic [38] for testing that the stratified treatment effect is null, which was originally derived for such conditional distributions, is unaffected by such tables.

Next, consider the random effects approach. Since it borrows from the whole, one obtains a finite estimate of a_k even when $s_k = 0$ or $f_k = 0$. Strata with $s_k = 0$ or $f_k = 0$ are relevant for the random effects model (4) also in terms of estimating the variance σ^2 of the centre estimates.

Deleting such a stratum usually has a decreasing effect on $\hat{\sigma}$, since the remaining strata show less variability in their overall success rates. Certainly, one would want to utilize data from all the centres if one were interested in estimating centre variability or individual centre effects $\{a_k\}$. Normally such tables have little effect on $\hat{\beta}$ or inference about β, an exception being mentioned below. Similar comments apply to random effects models for the relative risk or difference of proportions.

For inference with the relative risk or the difference in proportions, we next study analyses based on M–H estimators, for which effects of 0 column totals in strata are clear from the relevant formulae. The relative risk estimator (8) is unaffected by strata with $s_k = 0$, but strata with $f_k = 0$ provide a shrinkage toward 1.0. This is sensible, since when the two sample proportions of success fall within a small $\varepsilon > 0$ of 0, the sample relative risk can be any non-negative value, but when the two sample proportions are within ε of 1, the sample relative risk must fall very close to 1. Similarly, strata with $s_k = 0$ make no contribution to the estimated variance (10), and strata with $f_k = 0$ contribute to the denominator alone, thus providing a shrinkage in the variance estimate. For testing, strata with $s_k = 0$ make no contribution to the ratio of estimate to standard error, and provide no information about whether this type of effect exists.

For the M–H estimator (9) of the difference of proportions, strata with $s_k = 0$ or $f_k = 0$ make no contribution to the numerator but do contribute to the denominator. Thus, including such strata has the effect of shrinking the estimated difference of proportions toward 0 compared to the estimate that excludes them. This is expected, since such strata have a sample difference of proportions of 0. There is a compensating shrinkage effect on the standard error, and the ratio of estimate to standard error is unaffected by such strata. Thus, these strata also provide no information about whether this type of effect exists, although they do contribute toward estimating the size of the effect and hence provide evidence about whether interaction exists.

6.2. *Analyses of Table V*

Keeping in mind the highly tentative nature of any random effects modelling with such a small K, we summarize in Tables VI and VII various logit model analyses of Table V. The first row of Table VI reports estimates of the log-odds ratio β and their standard errors, for the ML fixed effects approach, the ML random effects approaches, and the M–H approach. Results are similar for all approaches, with the estimated common log-odds ratio of 1.5 (odds ratio of about 4.5) being about 2.2 standard errors.

The two centres with no successes can provide no information about the log-odds ratio treatment effect β as estimated by the fixed effects model or the M–H method. Very similar results occur with the random effects approach for the reduced data set deleting centres 1 and 3, as shown in the second line of Table VI.

For the no interaction models, the first row of Table VII reports ML estimates of $\{\alpha_k\}$ for the fixed effects model (1) and approximate BLUP estimates of $\{a_k\}$ for the random effects model (4). Because $s_1 = s_3 = 0, \hat{\alpha}_1 = \hat{\alpha}_3 = -\infty$ for model (1). Software may provide misleading indications in such situations, and a danger sign is when standard errors are enormous compared to the estimates, reflecting the very flat log-likelihood. The values in Table VII are those reported by PROC GENMOD in SAS (Version 7). PROC LOGISTIC provides $\hat{\alpha}_1 = -15.0$ (standard error $= 312.8$) and $\hat{\alpha}_3 = -15.3$ (standard error $= 339.7$) but warns that the ML estimates may not exist. The other centre estimates are the same for both procedures and the same as one obtains by deleting centres 1 and 3 from the data set (see row 2 of Table VII).

Table VI. Estimated treatment log-odds ratio (standard error in parentheses) for various logit models with Table V.

Data	Logit model (equation number)				
	Fixed, no interaction (1)	Random, no interaction (4)	Random non-parametric, no interaction	Random interaction (6)	Mantel–Haenszel (7)
Table V unadjusted	1.55 (0.70)	1.52 (0.70)	1.53 (0.69)	1.52 (0.70)	1.55 (0.71)
Delete centres 1,3	1.55 (0.70)	1.48 (0.70)	1.51 (0.69)	1.48 (0.70)	1.55 (0.71)
Combine centres 1–3	1.56 (0.70)	1.54 (0.70)	1.53 (0.69)	1.54 (0.70)	1.56 (0.70)
Add 0.000001 all cells	1.55 (0.70)	1.52 (0.70)	1.53 (0.69)	1.52 (0.70)	1.55 (0.71)
Add 0.05 all cells	1.48 (0.68)	1.45 (0.67)	1.46 (0.67)	1.45 (0.67)	1.48 (0.68)

Corresponding odds ratio estimates vary between $e^{1.45} = 4.3$ and $e^{1.56} = 4.8$.

As noted before, naive standard errors of estimates of random effects ignore the fact that the variance of those random effects is itself estimated. (Moreover, one is naive to expect to estimate well a variance component when K and n are as small as in the examples of this article!) Booth and Hobert [42] proposed a method for calculating standard errors based on the conditional mean squared error of prediction (CMSEP), given the data. This method incorporates a positive correction for the variability of the parameter estimates as well as an estimate of the bias incurred by using an estimate for the unknown conditional variance. Although this bias is often larger than the variance correction and thus non-ignorable, it is computationally difficult to calculate. Morris [41] proposed an analytic correction which can work well for the logistic mixed model [61]. Table VII reports the standard errors for the random effects centre estimates provided by NLMIXED, using the PREDICT option, which are based on a Laplace approximation to the CMSEP.

An ML estimate $\hat{\alpha}_k = -\infty$ is not very appealing when one truly believes that $\pi_{ik} > 0$. Because of the normality assumption, the random effects estimate of a_k also uses information from other centres and is finite. For centre 1, for instance, the estimate $\hat{a}_1 = -1.07$ provides an estimated success probability of $\exp(-1.07)/[1 + \exp(-1.07) = 0.255$ for placebo, even though that group had no successes at that centre. The estimated standard deviation of the centre effects is $\hat{\sigma} = 1.8$. Although centres with $\min(s_k, f_k) = 0$ provide no information about the treatment effect, deleting them from the analysis will tend to decrease $\hat{\sigma}$. In this case, $\hat{\sigma}$ decreases to 1.1.

The no interaction models, whether fixed effects or random effects, showed moderate evidence of a treatment effect. The random effects model permitting interaction has identical results (see Table VI), since the ML estimate of the standard deviation of the log-odds ratio is 0. This also happens when deleting centres 1 and 3 or when combining centres 1–3.

6.3. Extreme cases: centres with one observation per treatment

Simplified forms of the various estimates and standard errors occur for matched pairs data in which each row of each stratum contains a single observation. This is an extreme form of sparseness

Table VII. Estimated centre effects (standard error in parentheses) for no interaction models with Table V.

Data	Fixed effects model (1)					Random effects model (4)				
	α_1	α_2	α_3	α_4	α_5	a_1	a_2	a_3	a_4	a_5
Table V unadjusted	−28.0 (2.1×10^5)	−4.2 (1.2)	−27.9 (1.9×10^5)	−1.0 (0.7)	−2.0 (0.7)	−1.1 (1.4)	−0.5 (1.2)	−1.5 (1.4)	2.3 (1.2)	1.4 (1.2)
Delete centres 1,3		−4.2 (1.2)		−1.0 (0.7)	−2.0 (0.7)		−1.2 (0.9)		1.1 (0.8)	0.2 (0.8)
Combine centres 1–3	−4.9 (1.2)			−1.0 (0.7)	−2.0 (0.7)	−1.8 (1.1)			1.5 (1.0)	0.5 (1.0)
Add 0.000001 all cells	−16.6 (707.1)	−4.2 (1.2)	−16.8 (707.1)	−1.0 (0.7)	−2.0 (0.7)	−1.1 (1.4)	−0.5 (1.2)	−1.2 (1.4)	2.3 (1.2)	1.4 (1.2)
Add 0.05 all cells	−5.7 (3.2)	−4.1 (1.1)	−5.9 (3.2)	−0.9 (0.6)	−2.0 (0.6)	−0.9 (1.2)	−0.6 (1.0)	−1.0 (1.2)	2.1 (1.1)	1.2 (1.0)

Fixed effects estimates obtained using PROC GENMOD in SAS.

in which $n = 2K$. An important application is in cross-over studies, in which stratum k provides subject k's response for each treatment.

Let $a = \sum_k n_{11k} n_{21k}$ denote the number of pairs where both observations are successes, $b = \sum_k n_{11k} n_{22k}$ the number where the first is a success and the second is a failure, $c = \sum_k n_{12k} n_{21k}$ the number where the first is a failure and the second is a success, and $d = \sum_k n_{12k} n_{22k}$ the number where both are failures. Then, the M–H log-odds ratio estimate simplifies to

$$\hat{\beta}_{\mathrm{MH}} = \log(b/c), \quad \widehat{\mathrm{var}}(\hat{\beta}_{\mathrm{MH}}) = b^{-1} + c^{-1}$$

which is identical to the conditional ML estimate. Also, the M–H type of log relative risk estimate is

$$\hat{\phi}_{\mathrm{MH}} = \log[(a+b)/(a+c)], \quad \widehat{\mathrm{var}}(\hat{\phi}_{\mathrm{MH}}) = (b+c)/(a+b)(a+c)$$

and the M–H type of difference of proportions estimate is

$$\hat{\delta}_{\mathrm{MH}} = (b-c)/K, \quad \widehat{\mathrm{var}}(\hat{\delta}_{\mathrm{MH}}) = [(b+c) - (b-c)^2/K]/K^2$$

where $K = a + b + c + d$.

For this degree of sparseness, it is inappropriate to use ordinary ML estimators of these parameters based on models such as (1), as such estimators are inconsistent. The random effects version (4) is adequate, since the number of parameters in the marginal likelihood stays constant as K increases. In fact, suppose the association between the two responses is non-negative, in the sense that $\log(ad/bc) \geqslant 0$; then, for any parametric random effects model that is consistent with the data, the estimate of the log-odds ratio β is identical [62] to the M–H and conditional ML estimates, namely $\log(b/c)$.

6.4. *Effects of adding constants or combining centres*

When finite ML estimates do not exist, one approach in fixed effects models for contingency table analysis is to add a small positive constant to each cell (or to the empty cells), thus ensuring that all resulting estimates are finite. When that constant is small, however, the resulting value of $\hat{\beta}$ for model (1) and its standard error are usually almost identical to what one obtains by ignoring

strata for which $s_k = 0$ or $f_k = 0$. Table VI illustrates, showing the effect of adding 0.000001 to each cell and adding one observation to the data set (1/20 to each cell) for the sparse data of Table V. The treatment effects and goodness-of-fit are stable, as the addition of any such constant less than 0.001 to each cell yields $\hat{\beta} = 1.55$ (ASE = 0.70) and a G^2 goodness-of-fit statistic equal to 0.50.

Although this process also provides finite centre estimates for strata with $\min(s_k, f_k) = 0$, the estimates for these strata depend strongly on the constant chosen. Table VII illustrates, again showing the effect for added constants of 0.000001 and 0.05. The *ad hoc* nature of this approach is a severe disadvantage. Random effects and Bayesian approaches seem more suited to smoothing effects of zeros, and do not require adding arbitrary constants. Thus, we do not advocate adding constants in order to artificially include data from certain centres in the analysis.

An alternative strategy in multi-centre analyses combines centres of a similar type. Then, if each resulting partial table has responses with both outcomes, the ordinary descriptions and inferences use all the data. This, however, can affect somewhat the interpretations and conclusions made from those inferences. An extreme form of combining centres results from adding together all K tables and performing inference and description for that marginal X-by-Y 2×2 table. Although apparently sometimes done in practice, this can be dangerous, as Simpson's paradox [60] illustrates.

It seems reasonable to combine two centres if the descriptive measure of interest is similar for each and similar to what one gets by combining them. Sufficient conditions exist for when this happens. For instance, suppose the relative risk or the difference of proportions is identical for two centres. Then, the value of that measure takes the same value when the centres are combined [63] if the sample size ratio n_{1+k}/n_{2+k} is the same for each centre. For the odds ratio, collapsibility is more complex, sufficient conditions being the conditional independence of Z with either X or Y for those two strata. These conditions have limited relevance here, however, since when $\min(s_k, f_k) = 0$, there is no information about the size of the relative risk or odds ratio for that centre. Thus, it seems dangerous to combine that centre with others unless there are good reasons to believe that those centres are very similar and could be expected to share similar values of the measure of interest. For the difference of proportions, it is unnecessary in any case to combine a centre having $\min(s_k, f_k) = 0$ with other centres, since it makes a contribution as it is to the summary difference of proportions (although, as noted above, it provides no information about the significance of that difference).

For Table V, perhaps centres 1 and 3 are similar to centre 2, since the success rate is also very low for that centre. Table VI also shows the results of combining these three centres and re-fitting the models to this table and the tables for the other two centres. Here, the effect is negligible. In summary, with frequentist approaches there seems to be no loss of information regarding the significance of treatment effects by simply deleting centres having $\min(s_k, f_k) = 0$, although it is useful to include them for random effects analyses designed to estimate centre variability.

6.5. *Assumptions in models*

For severely sparse data, effects of model misspecification can be especially worrisome. Rarely would it be possible to check assumptions about homogeneity of effects or about a form of distribution for random effects. In estimating parameters in random effects models with sparse data, one might be concerned about how much those estimates may depend on the assumption for the random effects distribution. One way to check this assumption is to compare results to those obtained with a distribution-free approach for the random effects distribution, [64, 65] which

estimates that distribution using a finite number of mass points and probabilities. This approach is available with a GLIM macro [66]. Results for examples here are very similar to those assuming a normal random effect. Table VI illustrates for the no interaction model applied to Table V, providing results supplied by that GLIM macro.

At a minimum, it seems sensible to conduct some analyses designed to investigate sensitivity to assumptions and the influence of changes in the model and slight changes in the data. A *model sensitivity* study checks whether conclusions about the treatment effect are similar for a variety of plausible models. A *case sensitivity* analysis checks the effect on estimates and test statistics of deleting or adding a single observation or changing a single observation from success to failure or vice versa, checking this separately for each cell in the contingency table.

We illustrate the case sensitivity analysis for the no interaction random effects model with Table V. Checking the influence of each observation by deleting it from the data set, the estimated mean log-odds ratios vary from 1.42 to 1.87 with standard error ranging from 0.69 to 0.77, compared to the values of 1.52 and 0.696 for the observed data; the ratios of estimates to standard errors range from 2.02 to 2.42, compared to the observed $1.519/0.696 = 2.18$. The two smallest estimates and ratios result from deleting a success for the active drug in centre 4 or 5. When we instead add a single observation, the estimated mean log-odds ratios range from 1.16 to 1.61 with standard errors ranging from 0.63 to 0.70, while the ratios of estimates to standard errors range from 1.84 to 2.36. The five smallest estimates and ratios result from adding a success in the placebo group, in turn for each centre. After changing a single observation from success to failure or vice versa, the estimated mean log-odds ratios vary from 1.15 to 1.90 with standard errors ranging from 0.63 to 0.77; the ratios of estimates to standard errors range from 1.81 to 2.47. These results indicate the very tentative nature of any conclusions about the significance of the results in Table V.

As mentioned, it can be difficult to estimate well the variance components or standard errors of those components or the random effects. To check whether certain ones seem plausible, one might use the jack-knife or else treat the fitted model as if it were the true one and conduct a parametric bootstrap for independent binomials of the given row sizes satisfying that model [59]. This may be useful also to provide alternative confidence intervals. There is no guarantee that bootstrap methods will work well for highly sparse data, but a dramatically different result can suggest potential problems with the standard error estimate and corresponding Wald interval estimates.

6.6. *Dependence of results on method of fitting*

When using Gauss–Hermite quadrature to approximate the likelihood function in obtaining ML estimates for random effects models, the resulting quality of the approximations for the ML estimates can depend strongly on the number of quadrature points used. This is especially true when the data are sparse or the variance components are large. We recommend that the number of quadrature points be increased until the change in parameter estimates and standard errors is negligible. In our experience the standard errors and variance component estimates usually require a greater number of quadrature points for convergence than the treatment parameter estimates.

The number of quadrature points can be greatly decreased by centring the quadrature nodes at the mode of the function being integrated and scaling them by the curvature at the mode [24, 25]. Using this approach with the no interaction model for Table I, we needed only 9 quadrature points to obtain convergence (to four decimal places) in the parameter estimates and about 13 for the

standard errors, as opposed to about 200 quadrature points (about 270 for the standard errors) using the standard Gauss–Hermite nodes and weights. With the centred nodes approach one must take care when calculating predicted centre effects and interaction effects, since the functions being approximated may not be unimodal.

By default, PROC NLMIXED in SAS selects the number of quadrature points. Starting with one quadrature point, the log-likelihood is evaluated at the parameter starting values. The number of quadrature points is then increased and the log-likelihood re-evaluated until the difference between two successive evaluations is less than some user-controlled epsilon. That necessary number of quadrature points at the initial values is then used in all successive cycles in determining the parameter values that maximize the likelihood function. In our experience this often leads to only five or six points and can be inadequate for standard error calculations or predictions. Users can avoid the default method by using the QPOINTS= option. We also recommend expressing the variance components as products of standard deviations in the RANDOM statement of NLMIXED. Estimation of the standard deviation often avoids convergence problems when the estimated variance component is close to zero.

7. SUMMARY COMMENTS AND RECOMMENDATIONS

7.1. Similarities and differences in substantive results

For the examples in this paper, we reached similar conclusions about the treatment effect whether we used fixed effects or random effects models. Our experience with a variety of examples indicates that the fixed effects model and the random effects model assuming no interaction tend to provide similar results about the common treatment effect. Those results are also similar to the ones for the mean of the treatment effects for the random effects interaction model when the variance component estimate for the treatment effects equals 0 or close to 0. The latter model may provide a much wider confidence interval for the average effect when that variance component estimate is substantial. To illustrate, we alter Table I slightly, changing three of the failures to successes for drug in centre 3 and three of the successes to failures for drug in centre 8. Then the ML estimates are $\hat{\beta} = 0.759$ with $\mathrm{SE} = 0.305$ for fixed effects model (1) and $\hat{\beta} = 0.722$ with $\mathrm{SE} = 0.299$ for the random effects model (4) without interaction, but $\hat{\beta} = 0.767$ with $\mathrm{SE} = 0.623$ for the random effects model (6) permitting interaction. For the latter model, $\hat{\sigma}_b = 1.37$, compared to 0.15 for the actual data.

By contrast to the usual similarity of estimates of overall treatment effects with fixed effects and random effects models, the two model types can provide quite different estimates of individual centre or treatment effects. For instance, when all observations in a centre fall in the same outcome category, the random effects models smooth the centre effects considerably from the infinite values obtained with the fixed effects models.

An interesting question is to study the types of sparse data configurations or highly unbalanced data sets that can result in the two types of analyses giving substantively different treatment estimates or inferences about the treatment effect. As an extreme example (mentioned to us by Dr C. McCulloch in a personal communication), suppose that some strata have observations only for treatment 1 and the other strata have observations only for treatment 2. The fixed effects approach has insufficient information to estimate the treatment effect, since there are K binomial observations but $K + 1$ parameters for the no interaction model. By contrast, with the random effects approach

data of this form provide information about the treatment effect and about a mean and standard deviation of the random effects distribution, at least if one can regard the strata of each of these two types (having observations with only treatment 1 or having observations with only treatment 2) as a random sample from that distribution.

7.2. Strategies for choice of model and analysis

In selecting a method, a key determinant is the intended scope of inferences. If the strata truly are a sample of all possible strata and one would like to make inferences that apply more generally than to only the strata sampled, then the random effects approach is more natural. Data from multi-centre clinical trials and meta analyses are usually of that type, although the samples are usually not random. However, many share the view quoted earlier of Grizzle [16] that a random effects approach still better reflects all the actual sources of variability. If the strata sampled are the only ones of interest, such as when the strata are levels of control variables such as gender and race, the fixed effects approach is natural. Even when the strata are not a sample, however, the random effects estimators can be beneficial because of their smoothing effects. For instance, when there is significant interaction, the random effects estimates of stratum-specific log-odds ratios might be preferred to the separate sample values, especially when some of those sample values are infinite. See Senn [14] for a more sceptical view noting potential problems with using random effects approaches.

The choice of a fixed effects or random effects analysis can be a complex one having many considerations. [14, 16]. Among statistical considerations, for random effects modelling one should preferably have many more centres than the 8 in Table I and the 5 in Table V, yet the combining of information that occurs with random effects modelling is often very appealing. Among non-statistical considerations, a 'centre' is often quite arbitrary and not as well defined as a 'subject', yet we develop treatments not just for the subjects who attended the centres used in the study [14]. A referee has pointed out that one could consider 'fixed' and 'random' as but two labels for a continuum of sampling models that includes, for instance, systematic cases that are more representative than a random sample in certain senses and illustrative cases that are less so. Further development of such a framework of types of effects would be an interesting topic for further research.

Next, whatever one's choice of fixed or random effects model, one must decide whether to include interaction terms in the model. With many strata or highly sparse data, the power of tests of the hypothesis of no interaction may be weak. The safest approach is then to use the interaction model; otherwise, if one uses the simpler model but interaction truly exists, the standard error of the estimated treatment effect may be unrealistically low. Fixed effects and random effects no interaction models will tend to report smaller standard errors for the treatment effect than the interaction model, since the latter model permits an extra component of variance. Even when $\hat{\sigma}_b = 0$, the likelihood function often reveals that values of σ_b quite far from 0 are also plausible; thus, it is safest to use the interaction model. One may pay a penalty for doing so, having an increased standard error, but this simply reflects scepticism about the homogeneity model and the desire for inferences to apply more generally than for only the centres sampled.

With the random effects approach, one must also consider the validity of a normal assumption for the random effects. When the primary interest is in the treatment effect, the choice of distribution for the random effect should not be crucial [67], as a wide variety of mixing distributions lead to similar marginal distributions (averaged over the random effects). For model (4), for instance, if the normal distribution can induce an intracluster correlation approximately equal to the intracluster

correlation for the actual mixing distribution, then there is little bias in estimation of β or in the standard error estimates [67]. When the actual distribution is highly skewed, some bias [65, 67] may occur in estimating α.

The above remarks refer to the treatment effect. When estimation of centre effects are the focus, it is of interest to study the degree to which the estimates could depend on the choice of distribution. For fixed K, asymptotically this does not seem to be a problem. For instance, when the additive model form (1) holds, for any finite set of centre effects, as $\{n_{ik}\}$ increase the random effect estimators of treatment and centre parameters behave like the fixed effect estimators; in particular, both sets converge to the true values. In practice this manifests itself by the random effects estimates being very similar to the fixed effects estimates when the stratum-specific sample sizes are large.

An interesting open question is to study the effect of misspecification of the random effects distribution for the sparse asymptotic framework in which K grows with n. It is then too much to ask for consistency of centre estimates, but does one obtain consistency of estimation of the treatment effect and the variance components? One way to check the effect of the normality assumption is to compare results to those obtained with a non-parametric approach [64]. An advantage of the normal choice, other than convenience, is that it extends naturally to multivariate random effects that may have some correlation structure.

We have seen that centres with 0 successes or 0 failures can be disregarded in terms of deciding whether a treatment effect exists. They are needed, however, for estimating the variance component of centre effects in the random effects model, and for estimating the size of the effect in fixed and random effects models for the difference of proportions.

7.3. Extensions and alternative methods

Our emphasis has been on binary data with two groups, but the models and issues discussed generalize to multinomial data and several groups. For instance, for an ordinal response, one can use a proportional odds model with centre and treatment effects, with the centre effects being treated either as fixed or random. Recent work has focused on ways of fitting such models with random effects [68, 69], and one can use NLMIXED to fit the proportional odds model and related models (such as with probit link) based on the Gauss–Hermite quadrature approximation of the likelihood function.

Finally, this paper has focused on frequentist approaches. Alternative approaches include Bayes and empirical Bayes methods [4, 6, 11, 12, 70, 71]. The random effects model has much in common with empirical Bayes, in that it assumes a distribution for a set of parameters and uses the data to estimate parameters of that distribution.

ACKNOWLEDGEMENTS

This work was partially supported by grants from NIH and NSF. We appreciate helpful comments from Brent A. Coull, Ranjini Natarajan, Ramon Littell, Brett Presnell, Russell Wolfinger, and three referees.

REFERENCES

1. Beitler PJ, Landis JR. A mixed-effects model for categorical data (correction **42**: 1009). *Biometrics* 1985; **41**:991–1000.
2. Draper D, Gaver Jr, DP, Goel PK, Greenhouse JB, Hedges LV, Morris CN, Tucker JR, Waternaux CMA, Berlin JAR. *Combining Information*: *Statistical Issues and Opportunities for Research*. National Academy Press: 1992.
3. Lipsitz SR, Dear KBG, Laird NM, Molenberghs G. Tests for homogeneity of the risk difference when data are sparse. *Biometrics* 1998; **54**:148–160.

4. Berry SM. Understanding and testing for heterogeneity across 2×2 tables: application to meta-analysis. *Statistics in Medicine* 1998; **17**:2353–2369.
5. Givens GH, Smith DD, Tweedie RL. Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science* 1997; **12**:221–250.
6. Carlin JB. Meta-analysis for 2×2 tables: a Bayesian approach. *Statistics in Medicine* 1992; **11**:141–159.
7. Normand S-LT. Meta analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine* 1999; **18**:321–359.
8. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
9. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. *Statistics in Medicine* 1995; **14**:395–411.
10. Emerson JD. Combining estimates of the odds ratio: the state of the art. *Statistical Methods in Medical Research* 1994; **3**:157–178.
11. Skene AM, Wakefield JC. Hierarchical models for multicentre binary response studies. *Statistics in Medicine* 1990; **9**:919–929.
12. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* 1995; **14**:2685–2699.
13. Gallo PP. Practical issues in linear models analyses in multicentre clinical trials. *Biopharmaceutical Report of the Biopharmaceutical Section of the American Statistical Association* 1998; **6**:1–9.
14. Senn S. Some controversies in planning and analysing multi-centre trials. *Statistics in Medicine* 1998; **17**:1753–1765.
15. Jones B, Teather D, Wang J, Lewis JA. A comparison of various estimators of a treatment difference for a multi-centre clinical trial. *Statistics in Medicine* 1998; **17**:1767–1777.
16. Grizzle JE. Letter to the editor. *Controlled Clinical Trials* 1987; **8**:392–393.
17. Pierce DA, Sands BR. Extra-Bernoulli variation in regression of binary data. Oregon State University, Department of Statistics Technical Report 46, 1975.
18. Breslow N. Regression analysis of the log odds ratio: a method for retrospective studies. *Biometrics* 1976; **32**:409–416.
19. Hanfelt JJ, Liang K-Y. Inference for odds ratio regression models with sparse dependent data. *Biometrics* 1998; **54**:136–147.
20. Davis LJ. Generalization of the Mantel-Haenszel estimator to nonconstant odds ratios. *Biometrics* 1985; **41**:487–495.
21. Platt R, Leroux B, Breslow N. Generalized linear mixed models for meta-analysis. *Statistics in Medicine* 1999; **18**:643–654.
22. Coull BA, Agresti A. Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics* 2000; **56**:162–168.
23. Anderson DA, Aitkin M. Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society*, *Series B*, *Methodological* 1985; **47**:203–210.
24. Liu Q, Pierce DA. A note on Gauss–Hermite quadrature. *Biometrika* 1994; **81**:624–629.
25. Wolfinger RD. Towards practical application of generalized linear mixed models. Proceedings of 13th International Workshop on Statistical Modeling. Marx B, Friedl H. (eds), 1998:388–395.
26. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; **88**:9–25.
27. Wolfinger R, O'Connell M. Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* 1993; **48**:233–243.
28. McCulloch C. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* 1997; **92**:162–170.
29. Booth JG, Hobert JP. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society*, *Series B* 1999; **61**:265–285.
30. Zeger SL, Karim MR. Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* 1991; **86**:79–86.
31. Hobert J, Casella G. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* 1996; **91**:1461–1473.
32. Natarajan R, McCulloch CE. A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika* 1995; **82**:639–643.
33. Ghosh M, Ghosh A, Chen M, Agresti A. Noninformative priors for one parameter item response models, *Journal of Statistical Planning and Inference* 2000: in press.
34. Greenland S, Robins JM. Estimation of a common effect parameter from sparse follow-up data (correction **45**: 1323–1324). *Biometrics* 1985; **41**:55–68.
35. Andersen EB. Discrete *Statistical Models with Social Science Applications.* North-Holland/Elsevier: New York, 1980.
36. Breslow N, Day NE. *Statistical Methods in Cancer Research, Vol I: The Analysis of Case-Control Studies.* IARC: Lyon, 1980.

37. LogXact. *Logistic Regression Software Featuring Exact Methods.* Cytel Software: Cambridge, MA, 1993.
38. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 1959; **22**:719–748.
39. Tarone RE. On summary estimators of relative risk. *Journal of Chronic Diseases* 1981; **34**:463–468.
40. Nurminen M. Asymptotic efficiency of general noniterative estimators of common relative risk. *Biometrika* 1981; **68**:525–530.
41. Morris CN. Parametric empirical Bayes inference: theory and applications (correction pp. 55–65). *Journal of the American Statistical Association* 1983; **78**:47–55.
42. Booth JG, Hobert JP. Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association* 1998; **93**:262–272.
43. Breslow NE, Lin X. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* 1995; **82**:81–91.
44. Liu Q, Pierce DA. Heterogeneity in Mantel-Haenszel-type models. *Biometrika* 1993; **80**:543–556.
45. Raghunathan TE, Ii Y. Analysis of binary data from a multicentre clinical trial. *Biometrika* 1993; **80**:127–139.
46. Liang K-Y, Self SG. Tests for homogeneity of odds ratios when the data are sparse. *Biometrika* 1985; **72**:353–358.
47. Littell RC, Milliken GA, Stroup WW, Wolfinger RD. *SAS System for Mixed Models*. SAS Institute Inc.: Cary, NC, 1996.
48. Gart JJ, Nam J. Approximate interval estimation of the difference in binomial parameters: Correction for skewness and extension to multiple tables (correction **47**:357; **47**:979). *Biometrics* 1990; **46**:637–643.
49. Robins J, Breslow N, Greenland S. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* 1986; **42**:311–323.
50. Sato T. Comments on 'Estimation of a common effect parameter from sparse follow-up data' (**41**:55–68). *Biometrics* 1989; **45**:1323–1324.
51. Tarone RE. On heterogeneity tests based on efficient scores. *Biometrika* 1985; **72**:91–95.
52. Zelen M. The analysis of several 2×2 contingency tables. *Biometrika* 1971; **58**:129–137.
53. StatXact. *A Statistical Package for Exact Nonparametric Inference* (*version 4.0*). Cytel Software: Cambridge, MA, 1998.
54. Cochran WG. Some methods for strengthening the common chi-square tests. *Biometrics* 1954; **10**: 417–451.
55. Gart JJ. On the combination of relative risks. *Biometrics* 1962; **18**:601–610.
56. Radhakrishna S. Combination of results from several 2×2 contingency tables. *Biometrics* 1965; **21**: 86–98.
57. Emerson JD, Hoaglin DC, Mosteller F. A modified random-effect procedure for combining risk difference in sets of 2×2 tables from clinical trials. *Journal of the Italian Statistical Society* 1993; **2**:269–290.
58. Morris C. Central limit theorems for multinomial sums. *Annals of Statistics* 1975; **3**:165–188.
59. Simonoff JS. Jackknifing and bootstrapping goodness-of-fit statistics in sparse multinomials. *Journal of the American Statistical Association* 1986; **81**:1005–1011.
60. Agresti A. *An Introduction to Categorical Data Analysis*. Wiley: New York, 1996.
61. Greenland S. Second-stage least squares versus penalized quasi-likelihood for fitting hierarchical models. *Statistics in Medicine* 1997; **16**:515–526.
62. Neuhaus JM, Kalbfleisch JD, Hauck WW. Conditions for consistent estimation in mixed-effects models for binary matched-pairs data. *Canadian Journal of Statistics* 1994; **22**:139–148.
63. Shapiro SH. Collapsing contingency tables — A geometric approach. *American Statistician* 1982; **36**: 43–46.
64. Aitkin M. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* 1999; **55**:117–128.
65. Heckman J, Singer B. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 1984; **52**:271–320.
66. Aitkin M, Francis B. Fitting overdispersed generalized linear models by non-parametric maximum likelihood. *GLIM Newsletter* 1995; **25**:37–45.
67. Neuhaus JM, Hauck WW, Kalbfleisch JD. The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* 1992; **79**:755–762.
68. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics* 1994; **50**: 933–944.
69. Tutz G, Hennevogl W. Random effects in ordinal regression models. *Computational Statistics & Data Analysis* 1996; **22**:537–557.
70. Liao JG. A hierarchical Bayesian model for combining multiple 2×2 tables using conditional likelihoods. *Biometrics* 1999; **55**:21–26.
71. Louis TA. Using empirical Bayes methods in biopharmaceutical research (Discussion pp. 828–829). *Statistics in Medicine* 1991; **10**:811–827.

TUTORIAL IN BIOSTATISTICS

A REVIEW OF TESTS FOR DETECTING A MONOTONE DOSE–RESPONSE RELATIONSHIP WITH ORDINAL RESPONSE DATA

CHRISTY CHUANG-STEIN[1]* AND ALAN AGRESTI[2]

[1] *Clinical Development Biostatistics, Pharmacia & Upjohn Company, Kalamazoo, MI 49001, U.S.A.*
[2] *Department of Statistics, University of Florida, Gainesville, Florida 32611-8545, U.S.A.*

SUMMARY

This tutorial reviews methods for testing independence between discrete levels of a dose and an ordered categorical response variable. The tests are designed to be powerful for cases in which the response improves monotonically as dosage level increases. First, we show how to apply some standard tests for doubly-ordered contingency tables. Then, we show how to construct tests as part of a model-building strategy. Other topics discussed include generalizations to stratified data, small-sample methods, and sample size and power considerations. © 1997 by John Wiley & Sons, Ltd.

Statist. Med., **16**, 2599–2618 (1997)
No. of Figures: 0 No. of Tables: 5 No. of References: 50

1. INTRODUCTION

The exploration of dose–response relationships is the focus of many studies in toxicology[1] and genetic toxicology.[2] This topic occupies an equally important place in animal growth promotion studies[3] and in the pre-marketing clinical testing of new drugs. In a typical pre-marketing dose–response study, a control and several doses of the drug are randomly assigned to the study subjects, with each subject receiving only one dose throughout the study (this is called the parallel-group design). The control is usually a placebo that provides necessary background information.

In dose–response studies, the response can either measure the efficacy of a treatment or the risk (side-effect) associated with an exposure. The exposure might be that to a new medication or to a risk factor such as cigarette smoking, with the dose quantifying the amount of exposure.

Statisticians have used the phrase *dose–response relationship* to represent a variety of things. Some refer to the shape of the exposure–outcome curve, no matter what that shape may be.[4] For

* Correspondence to: Christy Chuang-Stein, Clinical Development Biostatistics, Pharmacia & Upjohn Company, Kalamazoo, MI 49001, U.S.A. E-mail: jcchuang@au.pnu.com.

Tutorials in Biostatistics Volume 1: Statistical Methods in Clinical Studies Edited by R. B. D'Agostino
 ISBN: 0-470-02365-1

Table I. Responses on the Glasgow Outcome Scale from a clinical trial with a placebo (control) and three treatment groups labelled as low dose, medium dose and high dose

Treatment group	Glasgow Outcome Scale					Total
	Death	Vegetative state	Major disability	Minor disability	Good recovery	
Placebo	59	25	46	48	32	210
Low dose	48	21	44	47	30	190
Medium dose	44	14	54	64	31	207
High dose	43	4	49	58	41	195

some, the objectives of a study assessing exposure-associated risk are to demonstrate a continuously increasing risk with increasing exposure.[5] Recently, other shapes have received attention, such as the umbrella pattern[6-10] and a plateaued drug effect beyond a certain level. Among the potential shapes, the monotone one is by far the most commonly discussed in the literature.

A rich literature exists on the exploration of dose–response relationships for the parallel-group design. The literature refers almost exclusively, however, to normal or binary response variables. The main purpose of this article is to summarize methods one can apply for ordinal responses – that is, responses measured with a set of ordered categories. Most of these methods were originally proposed for other applications, but are appropriate for dose–response relationships. The article assumes that the reader has familiarity with basic ideas of statistical inference, regression and ANOVA modelling, and chi-squared tests. Section 4 also assumes previous exposure to logit modelling, but otherwise the article does not require previous background in specialized methods for categorical response data.

Ruberg[11,12] noted that dose–response studies routinely ask four questions: (i) Is there any evidence of a drug effect? (ii) Which doses exhibit a response different from the control response? (iii) What is the nature of the dose–response relationship? (iv) Which is the optimal dose? One approaches the questions in this order, the later ones being more specific. In the drug development process, information obtained from dose–response studies is often used to select doses for subsequent confirmatory registration trials.

This article primarily focuses on the first question. Clearly, though, answers to the other questions are ultimately more informative. Though we occasionally refer to them as well, for lack of space we defer a detailed account to a follow-up paper. We summarize methods designed to detect an effect on an ordinal response when there is prior belief of a monotone dose–response relationship, expressed in the vague notion that a higher dose tends to produce a more desirable outcome. This prior belief relates to a monotone alternative to the 'no effect' hypothesis. We present the methods in the context of efficacy evaluation, though they also apply to risk assessment with a reversal in the direction of association. Strict monotonicity is not required, and we use 'monotone' interchangeably with 'non-decreasing'. This includes cases having only a high-dose effect or a constant drug effect at all the non-zero dose levels.[13]

Table I illustrates the type of data considered in this article. In Table I, five ordered categories ranging from 'death' to 'good recovery' describe the clinical outcome of patients who experienced trauma. In the literature on critical care, these five categories are often called the Glasgow Outcome Scale (GOS). Table I includes four treatment groups, with a vehicle infusion serving as the control. The three intravenous doses for the investigational medication are labelled as low,

medium and high. The original data have been modified somewhat to protect the identity of the trial. One study objective was to determine whether a more favourable GOS outcome tends to occur as the dose increases. This example and others in this article deal with fixed doses determined prior to the studies. As a result, levels of dose are treated as fixed rather than random. This is natural, since a pharmaceutical sponsor needs to justify its choice of the recommended dose in the new drug application for the compound. Furthermore, the availability of dosing strengths is often influenced and limited by manufacturing considerations.

Though Table I is simply a two-way contingency table, standard tests of independence for two-way tables such as the Pearson chi-squared test are inappropriate for testing against an ordered alternative. Those tests treat both classifications as nominal scale (unordered). When a monotone trend truly exists, methods designed to detect it are more powerful than such nominal-level procedures.

One possible approach dichotomizes the response and employs methods for binary responses. This approach is reasonable if the response categories are clearly divided into desirable and undesirable groups. Otherwise, this approach suffers from the lack of a clear choice for the collapsing and, as we see in Section 6, a loss of information and power.

The organization of this article is as follows: Sections 2 and 3, the heart of the article, present several tests of independence between dose and an ordinal response that are sensitive to the alternative of a monotone relationship. Section 2 discusses non-model-based tests while Section 3 focuses on model-based inference. Section 4 mentions generalizations to handle stratification. Section 5 discusses small-sample and sparse-data inference, and Section 6 comments on sample size and power. The final section summarizes and provides recommendations for conducting such an analysis.

2. SIGNIFICANCE TESTS FOR A MONOTONE DOSE–RESPONSE RELATIONSHIP

This section reviews significance tests for detecting monotone dose–response relationships. Section 3 discusses related tests for models for the relationship. Non-model-based inference, though less informative, is often considered simpler from a regulatory perspective because the tests do not need to validate any modelling assumptions;[14] however, we shall note in Section 3 that some tests from this section are equivalent to tests for certain models. This section presents four approaches: (i) tests based on association measures, including generalized Cochran–Mantel–Haenszel procedures; (ii) an adaptation of the Jonckheere–Terpstra test; (iii) adaptations of methods for continuous responses, including order-restricted inference; and (iv) treating the response distributions as survival distributions.

Let I denote the number of treatment groups, and let J denote the number of categories of the response variable, which is denoted by Y. Let x_{ij} denote the number of individuals in the ith treatment group whose response falls in the jth category, let $n_i = \sum_j x_{ij}$ denote the number of subjects in that group and let $N = \sum n_i$ denote the total sample size. We treat the counts in separate rows as independent multinomial samples. We arrange the I treatment groups from the lowest ($i = 1$) to the highest dose group ($i = I$), with d_i representing the dose level for the ith group, and the response categories from the least favourable ($j = 1$) to the most favourable ($j = J$).

Let Y_i denote a response at dose i. Let $F_{ij} = P(Y_i \leqslant j)$. The null hypothesis of no difference among the I treatment groups is

$$H_0: F_{1j} = F_{2j} = \cdots = F_{Ij} \text{ for all } j. \tag{1}$$

One way to operationalize the alternative hypothesis of 'monotone dose-response relationship' is in terms of a monotone *stochastic ordering* among the I cumulative distributions. This means that

$$H_1: F_{1j} \geqslant F_{2j} \geqslant \cdots \geqslant F_{Ij} \text{ for all } j \tag{2}$$

with strict inequality for at least one j. Since higher response categories represent more favourable outcomes, this alternative implies a tendency for more favourable outcomes as the dose increases.

Since significance tests relate to particular hypotheses, they are confirmatory rather than exploratory in nature. The use of significance tests relates not to exploring the nature of the dose–response relationship, but rather to determining the probability of results at least as extreme as those observed in the direction of a monotone relationship, if the variables were truly independent. As a result, testing for a monotone relationship only makes sense when the pharmacology of the drug suggests that, within the safety limits, higher drug exposure results in efficacy that is at least as good as that at lower exposure. We suggest combining such formal analyses with informal checks of this prior belief, such as by plotting sample cumulative distributions. The modelling approaches in Section 3 have the advantage of a built-in goodness-of-fit check.

2.1. Tests based on association measures

Table I has ordered rows (the doses) and ordered columns (the ordinal response). The doses are quantitative, and one can treat the response scale as quantitative by assigning scores to the categories. Correlation-type association measures then summarize the linear component of the dose–response relationship. This strategy is reasonable if one expects roughly a linear trend on the chosen scales. Yates[15] presented a large-sample single-degree-of-freedom chi-squared test statistic based on this approach, essentially squaring the ratio of the sample correlation to its standard error. To form a P-value for the one-sided alternative of a positive trend, we use the signed square root of this statistic and refer to the right-hand tail probability from the standard normal curve. For binary responses, the closely related *Cochran–Armitage*[16] *test* is designed to detect a linear trend in a response proportion. Mantel[17] extended Yate's test to the stratified case.

A potential disadvantage of this strategy is the necessity of assigning scores. Normally, one would assign the actual dosage level or the log dose to the dose categories. Usually, the choice of response scores has little effect on the conclusion about whether an effect exists. It may have an effect, however, when the data are highly unbalanced, such as when some categories have many more observations than other categories.[18]

An alternative approach with correlation measures avoids the responsibility of selecting scores and uses the data to form them automatically. Specifically, one assigns ranks to the subjects and uses them as the category scores. For all subjects in a category, one assigns the average of the ranks that would apply for a complete ranking of the sample. These are called *midranks*. Let $x_{+j} = \sum_i x_{ij}$ denote the number of subjects in the sample who make response j. The midrank for category j equals

$$w_j = x_{+1} + \cdots + x_{+,j-1} + x_{+j}/2, \quad j = 1, \ldots, J.$$

The use of midrank scores for the responses and midrank scores for the drug doses yields a generalization of Spearman's rho for contingency tables with ordered categories.

The use of rank-based scores seems appealing, since one does not need to select arbitrary scores, but midrank scores do not necessarily provide distances between categories that correspond to a 'reasonable' metric.[18] In particular, for highly unbalanced response frequencies, adjacent categories having relatively few observations necessarily have similar midrank scores, even if they seem far apart in practical terms. For example, suppose few subjects fell in the first three categories on the scale (death, fair, good, very good, excellent); midranks then have similar scores for the categories 'death' and 'good'. It is usually better to use one's judgement by selecting scores that reflect perceived distances between categories. When uncertain about this choice, one should perform a sensitivity analysis, selecting two or three 'sensible' choices and checking that the conclusions are similar for each; for instance, for the scale just mentioned, one might compare results for scores (0, 1, 2, 3, 4), (0, 5, 7, 9, 10) and (0, 7, 8, 9, 10). Equally-spaced scores often provide a reasonable compromise when the category labels do not suggest any obvious choices, such as the response categories (worse, no change, better).

The test statistic for the fixed-score or rank-score correlation approach is a special case for a single table of a generalized Cochran–Mantel–Haenszel (CMH) statistic for testing conditional independence with several $I \times J$ contingency tables.[19] That chi-squared statistic, having d.f. = 1, summarizes correlation information between two ordinal variables, combined over several strata. For a single table such as Table I, it equals $(N - 1)\, r^2$, where r denotes the sample correlation for the chosen scores. It is available in SAS (PROC FREQ) as the 'non-zero correlation' test, generated using option CMH1 in that procedure; one can use either fixed scores selected by the user or midrank scores. Table II shows SAS code for analysing Table I using (i) scores (1, 2, 3, 4) for dose and (1, 2, 3, 4, 5) for outcome and (ii) midrank scores. The signed square root of this generalized CMH statistic, $M = \sqrt{(N - 1)}\, r$, is a standard normal test statistic that is sensitive to the direction of trend.

The correlation-based test applied to Table I has $M = 3{\cdot}10$ using any sets of equally-spaced scores for the rows and the columns and $M = 3{\cdot}07$ using midrank scores, both having one-sided P-values of 0·001. The response scores (0, 1, 6, 9, 10), which may reflect a more reasonable assessment of distances between outcome categories, yield $M = 3{\cdot}59 (P < 0{\cdot}001)$ when used in combination with equally-spaced row scores; the response scores 0, 0, 1, 3, 10, which give much more weight to the most favourable outcome, yield $M = 2{\cdot}35$ ($P = 0{\cdot}009$). Each statistic provides strong evidence against the hypothesis of identical response distributions at the various dose levels.

A similar association test strategy, but not requiring any scores, bases the test on a measure that strictly uses ordinal information. Examples include the generalizations of Kendall's tau for contingency tables that utilize the numbers C of concordant and D of discordant pairs in summarizing information about an ordinal trend (Agresti,[20] pp. 22, 34). The standard measures fall between -1 and $+1$, have expectations of zero under the null hypothesis, and have approximate large-sample normal distributions. One can form a z test statistic (that is, having a standard normal null distribution) by dividing any such measure by its large-sample standard error.

These measures describe the extent of monotonicity in the relationship, without focusing on a particular aspect of it, such as linearity. An example is *Goodman and Kruskal's gamma*, which is $(C - D)/(C + D)$. Gamma equals the difference between the proportion of concordant pairs and the proportion of discordant pairs, out of the untied pairs. *Somers' d*, which treats the variables asymmetrically, is the difference between these proportions out of those pairs of observations falling at different dose levels. For instance, Somers' d equals 1·0 if, for each dose level, every response at that dose level exceeds every response at every lower dose level.

Table II. Example of SAS code for performing various analyses with Table I

```
data cmh;
input dose outcome count @@;
group = 1;
cards;
1 1 59    1 2 25    1 3 46    1 4 48    1 5 32
2 1 48    2 2 21    2 3 44    2 4 47    2 5 30
3 1 44    3 2 14    3 3 54    3 4 64    3 5 31
4 1 43    4 2  4    4 3 49    4 4 58    4 5 41
;
proc freq; weight count; * CMH with scores entered in data;
  tables group * dose * outcome / cmh1;
proc freq; weight count; * CMH with mid-rank scores;
  tables group * dose * outcome / cmh1 scores = ridit;
proc freq; weight count; * association measures such as gamma;
  tables dose * outcome / measures;
proc catmod order = data; weight count; * mean response model;
  population dose;
  response 1 2 3 4 5; direct dose; * uses scores (1, 2, 3, 4, 5);
  model outcome = dose;
proc logistic; freq count; * proportional odds model (ML);
  model outcome = dose;
proc catmod; weight count; * proportional odds model (WLS);
  response clogits; direct dose;
  model outcome = _response_ dose;
proc catmod; weight count; * adjacent cat. logit model (WLS);
  response alogits; direct dose;
  model outcome = _response_ dose;
run;
```

Formulae for standard errors of the extensions of Kendall's tau are quite complex. The measures and their estimated standard errors are available in standard software, such as SAS (PROC FREQ), as illustrated in Table II. For Table I, gamma = 0·118 and has a standard error of 0·038, leading to test statistic $z = 3{\cdot}11$ and $P = 0{\cdot}001$; Somers' d provides similar results, its value of 0·092 having a standard error of 0·030.

As in other contexts, the non-null expected values of various score-based correlation measures or ordinal association measures depend on the distribution of subjects to the various dose levels; the measures tend to increase with greater dispersion in the dose values. The test statistics based on them provide a simple way of summarizing trend information, even though the sample measure may not be used to estimate a particular population parameter.

2.2. Jonckheere–Terpstra test

For any pair $a < b$ of doses, the midranks for response levels for the $2 \times J$ table formed from these two treatment groups equal

$$w_{(ab)j} = (x_{a1} + x_{b1}) + \cdots + (x_{a,j-1} + x_{b,j-1}) + (x_{aj} + x_{bj})/2, \; j = 1, \ldots, J.$$

The Jonckheere–Terpstra (JT) test[21] statistic sums the $I(I-1)/2$ one-sided Wilcoxon–Mann–Whitney statistics for comparing pairs of treatment groups, in the order given by the

doses. In other words, the test statistic is based on

$$JT = \sum_{b=2}^{I} \sum_{a=1}^{b-1} \sum_{j} \left(w_{(ab)j} x_{bj} - \frac{n_b(n_b + 1)}{2} \right).$$

For large samples, the standardized value $z = [JT - E(JT)]/[\text{var}(JT)]^{1/2}$ provides a test statistic. Again, the variance formula is complex (see StatXact,[22] p. 614). The StatXact software, which provides a great variety of small-sample and asymptotic analyses for categorical data, can conduct this test. The same comments apply to this strategy as to the rank-based association measure approach presented in the previous subsection. For Table I, $z = 3{\cdot}10$, having one-sided P-value of 0·001.

2.3. Tests treating the response as continuous

A common approach for analysing ordinal data is to assign scores to the response categories and use standard normal-theory methods, such as regression and analysis of variance. From our experience, treating ordinal data as continuous with constant variance can provide a useful approximation when the number of response categories is large, but may be inadequate when that number is less than five. At the highest dose or at the no-dose level, responses often fall mostly in one category, yet are more dispersed at other dose values. Though this can cause problems for model building, for instance with predicting means or cell probabilities, it is less problematic for significance testing. For testing with small samples in the two-sample case, Heeren and D'Agostino[23] showed that the actual level of the t-test may not exceed the nominal level by much, but it can be considerably less than the nominal level.

When predictors are categorical, one can account for non-constant response variance by basing regression parameter estimates and standard errors explicitly on multinomial rather than normal assumptions for the response distribution. See, for instance, the mean response model discussed by Grizzle *et al.*[24] and Agresti[20] (Section 9.6). A weighted least squares (WLS) solution is simple to implement for this method using SAS (PROC CATMOD), as illustrated in Table II. When the data do not display widely varying dispersion or when the model fits well, the two approaches (ordinary and weighted least squares) provide very similar results.

For Table I, using the dose scores, the regression t-test for a normal response has $t = 3{\cdot}12$ ($P = 0{\cdot}001$) for equally-spaced response and dose scores, and $t = 2{\cdot}35$ ($P = 0{\cdot}009$) for response scores (0, 0, 1, 3, 10). The corresponding results using the methodology of Grizzle *et al.* are $z = 3{\cdot}10$ ($P = 0{\cdot}001$) and $z = 2{\cdot}25$ ($P = 0{\cdot}012$). For response scores (1, 2, 3, 4, 5), the prediction equation for the mean response is 2·699 + 0·138 (dose) both using ordinary and weighted least squares. For the scores (0, 0, 1, 3, 10), they are 2·089 + 0·256 (dose) and 2·099 + 0·248 (dose).

This approach has the advantages of fully utilizing the inherent quantitative nature of the variables and directing the focus toward model-building rather than significance testing. A disadvantage, compared to models discussed in Section 3, is that conclusions disregard the categorical nature of the response scale. For instance, models that treat the response as categorical provide predicted probabilities of response in each category.

2.4. Order-restricted tests treating the response as continuous

For the monotone stochastic ordering alternative, the approximate approach of treating the ordinal response as normal with constant variance can also utilize methods developed for testing

equality of normal means against order-restricted alternatives. We now review some methods in this class.

Bartholomew[25,26] proposed one of the earliest order-restricted methods. Denote the true mean and the sample mean of the ith dose group by μ_i and $\bar{y}_i$. Assuming normality, one obtains the maximum likelihood (ML) estimates $\{\hat{\mu}_i\}$ subject to the constraint $\mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_I$ by constructing the finest possible partition $\{R_l\}$ of treatment groups $\{1, \ldots, I\}$ so that

$$\frac{\sum_{i \in R_l} n_i \bar{y}_i}{\sum_{i \in R_l} n_i}$$

is strictly increasing in ℓ. For all i in R_l, $\hat{\mu}_i$ are identical and equal a weighted sample mean. The solution of order-restricted mean estimates is called the *isotonic regression* of $\bar{y}_i$ with respect to the simple order on the row means $\{\mu_i\}$, with $\{n_i\}$ as the weights.[27] The partition is easily determined with the *pooling adjacent violators algorithm.*[27,28]

Denote the jth observation in the ith group by y_{ij}. When the population variance is unknown, Bartholomew[26] proposed the test statistic

$$\bar{E}^2 = \frac{\sum_i n_i(\hat{\mu}_i - \bar{y})^2}{\sum_i \sum_j (y_{ij} - \bar{y})^2} \tag{3}$$

where $\bar{y} = (\sum n_i \bar{y}_i)/N$ is the overall sample mean. The large-sample distribution of $\bar{E}^2$ is nonstandard, being the same as that of a weighted average of beta random variables. Relatively large $\bar{E}^2$ values provide evidence against the null hypothesis. Robertson *et al.*[28] (Chapter 2) discussed this test, and Brunden[29] prepared an SAS program that computes the weights and supplies critical values.

For the equally-spaced response scores (1, 2, 3, 4, 5), the sample means are 2·852 (placebo), 2·947 (low dose), 3·116 (medium dose) and 3·256 (high dose). The sample means satisfy the order restriction, and $\bar{E}^2 = 0{\cdot}0122$. The sample means for the second choice of scores (0, 0, 1, 3, 10) are 2·429, 2·553, 2·686 and 3·246, again satisfying the order restriction, and $\bar{E}^2 = 0{\cdot}0081$. The upper 5th and 1st percentiles of the null $\bar{E}^2$ distribution (Brunden[29]) are 0·0056 and 0·0096, respectively, so the P-value is less than 0·01 for the equally-spaced scores and less than 0·05 (about 0·025) for the unequally-spaced scores.

A small P-value for an order-restricted test suggests strong evidence against the null hypothesis, but just as with previously mentioned tests, this does not imply that monotone ordering holds perfectly in the population of interest. Small P-values can occur when the expected order is violated somewhat in the sample, but the test statistic would be sufficiently unusual under the null. To illustrate, consider Table III, showing responses (worse, same, slightly better, much better) to three doses (placebo, low and high) for a hypothetical sample of 123 subjects. The sample means under the equally-spaced response scores (1, 2, 3, 4) are 2·72 (placebo), 2·62 (low dose), and 3·05 (high dose), violating the order restriction. The isotonic regression of these sample means with respect to the increasing order on the row means provides mean estimates of 2·67 (placebo and low dose) and 3·05 (high dose), and $\bar{E}^2 = 0{\cdot}038$. The sample means under another choice of scores ($-3, 0, 2, 5$) are 1·63, 1·36 and 2·39, again requiring pooling adjacent violators to obtain order-restricted mean estimates, which are 1·49, 1·49 and 2·39, and for which $\bar{E}^2 = 0{\cdot}033$. The upper 5th and 1st percentiles of the null $\bar{E}^2$ distribution with three treatment groups (Brunden[29]) are 0·031 and 0·055, respectively, so the P-value is a bit less than 0·05 for both choices of scores. For either choice of scores, the mean for the high dose group

Table III. Example that violates order restriction but yields small P-value

Treatment group	Response category			
	Worse	Same	Slightly better	Much better
Placebo	3	15	12	10
Low dose	4	17	12	9
High dose	2	8	17	14

is sufficiently large that, if there were truly no effect, it would be unusual to obtain such a large test statistic value.

When all $\{n_i\}$ equal some constant n and the objective is to compare several doses with a zero-dose control, Williams[30] proposed the test statistic

$$\bar{t} = \frac{\hat{\mu}_I - \bar{y}_1}{\sqrt{(2s^2/n)}} \tag{4}$$

where s^2 is an unbiased estimator of error variance. Williams tabulated the distribution of this statistic and generalized it[31] when each treatment sample size is a constant multiple of the control group sample size. Williams noted that $\bar{t}$ has higher power than (3) when a constant drug effect occurs ($\mu_2 = \cdots = \mu_I$), and showed its role in sequential testing to determine the lowest dose at which evidence exists of a drug effect. In other cases, (3) and other statistics that incorporate more of the sample information are likely to perform better. Capizzi *et al.*[32] reported a simulation study that further compared these two procedures with an adjustment of a trend test proposed by Tukey *et al.*[33] They found that the adjusted trend test tends to be more powerful than the other procedures, although circumstances exist where either Bartholomew's[34] or Williams'[31] test appears superior.

For Table I, we equated n to the arithmetic mean of n_i, since the study intended to assign the same number of patients to each treatment group. Williams' test yields $\bar{t} = 2{\cdot}877$ for equally-spaced scores and 2·367 for the unequally-spaced scores. From Williams,[30] the upper 5th and 1st percentiles of the null distribution for $\bar{t}$ are 1·739 and 2·377. Thus, the P-value is less than 0·01 for equally-spaced scores and barely exceeds 0·01 for the unequally-spaced scores. As for Table III, with n_i set to 41, Williams' test yields $\bar{t} = 1{\cdot}881$ for the equally-spaced scores and 1·748 for the unequally-spaced scores. The upper 5th and 1st percentiles of the null distribution for the case of three treatment groups are 1·731 and 2·400, so the P-values are similar to those obtained with the $\bar{E}^2$ statistic for this example.

Shirley[35] proposed a Wilcoxon-type version of Williams' test, with emphasis on comparing increasing doses of a substance with a zero-dose control. Hothorn[1] studied the robustness of Williams' and Shirley's[35] procedures as applied in toxicology studies. He concluded that Shirley's procedure tends to behave better when assumptions underlying the analysis of variance are violated. For additional discussion of order-restricted inference, see Robertson *et al.*,[28] Cohen and Sackrowitz,[36] Hayter,[37] Silvapulle and Silvapulle,[38] and the references therein.

For an ordinal response, the approaches just discussed are somewhat unsatisfactory, since they treat the response as normal with constant variance rather than multinomial. One might prefer an

order-restricted test derived specifically for the monotone stochastic ordering alternative, under the assumption of multinomial sampling. For $I = 2$, Grove[39] and Robertson and Wright[40] proposed a likelihood-ratio test for testing whether two multinomial distributions are identical against the alternative of a stochastic ordering. The large-sample distribution of the test statistic is chi-bar squared, the distribution of a weighted average of chi-squared variates with differing degrees of freedom. For an observed test statistic value t, the P-value has the form $\sum_{j=1}^{J} p_j \Pr(\chi^2_{j-1} > t)$, where $\{p_j\}$ are weights that are recursively calculated.

For $I > 2$, results for order-restricted comparisons of multinomial distributions are incomplete. Patefield[41] suggested using an alternative that is a special case of a stochastic ordering, but noted the computational complexity of maximizing the likelihood. Grove[42] proposed a large-sample chi-bar-squared test for a different type of ordered alternative. Tests for the ordinary stochastic ordering alternative for the several groups case do not seem to appear in the literature, but one such test is discussed in a recent technical report by Agresti and Coull (University of Florida, 1996).

2.5. Treating the response distributions as survival distributions

When response categories are ordered, such as in Table I, one can use methods that apply to life tables. To do this, one orders the response categories from the least to the most favourable ones and regards observations in column j like failures between times $j - 1$ and j, the most favourable category (for example, good recovery in Table I) representing subjects with censored lifetimes beyond time $J - 1$. In this formulation, $R_{ij} = \sum_{b \geq j} x_{ib}$ represents the number of subjects at risk prior to time j, for dose i. The approaches discussed in this subsection have the advantage of not requiring response scores.

For life-table analysis, Tarone[43] proposed a test for a trend in hazard functions as the dose level increases. The square of this trend statistic is a special case of the summary chi-squared statistic proposed by Mantel[17] for stratified tables with ordered levels. In this context, we express the data as $(J - 1)$ separate $I \times 2$ contingency tables, where the jth one compares the I doses on a binary response in which the first level is category j of the original response and the second level combines responses in all categories higher than j. The $(J - 1)$ component tables in this construction are independent, because the corresponding sets of 'continuation-ratio' binomial variates in the component tables are independent. The statistic uses the dose scores by summarizing the correlation between dose and this binary response across the $(J - 1)$ ways of forming the binary response. One can compute this correlation-type statistic by applying the CMH1 option in PROC FREQ in SAS to the $(J - 1)$ tables. For Table I, the chi-squared statistic equals 8·350 with d.f. = 1, with a one-sided P-value (for its positive square root) of 0·002.

For this statistic, reversing the order of the response categories yields a different value of the test statistic. For instance, applying the test in the reverse order to Table I, the chi-squared test statistic equals 7·08, giving a one-sided P-value of 0·004. This behaviour is not true for other tests discussed in this article, which have the same result for each of the two possible orders of categories for the ordinal scale.

3. MODEL-BASED INFERENCE ABOUT MONOTONE DOSE–RESPONSE RELATIONS

The tests in Section 2 are fine for detecting evidence against the null hypothesis in the direction of a positive trend. However, they do not lend much insight about the form of the relationship.

A model-based perspective is superior for this purpose. A good-fitting model describes the nature of the association, provides parameters for describing the strength of the relationship, provides predicted probabilities for the response categories at any dose, and helps us to determine the optimal dose. As a by-product, it also yields tests for the hypothesis of no effect. In fact, some tests presented in the previous section have natural connections with models. In this section, we again focus on the first question posed by Ruberg[11] – that is, whether an effect exists; however, the model-based approach is also well-suited for pursuing the other three questions.

The models we discuss are generalizations of logistic regression models that handle ordinal response categories. For further discussion of these and other models for ordinal responses, see Agresti[20] (Chapters 8 and 9) and McCullagh.[44]

3.1. Proportional odds models

Currently, the most popular model for ordinal responses uses logits of cumulative probabilities. A J-category response has $(J-1)$ non-redundant cumulative probabilities, $P(Y_i \leqslant j)$, $j = 1, \ldots, J-1$. For the dose–response problem, consider the model

$$\text{logit}[P(Y_i \leqslant j)] = \alpha_j - \beta_i, \quad i = 1, \ldots, I, \; j = 1, \ldots, J-1 \tag{5}$$

where $\text{logit}[P(Y_i \leqslant j)] = \log[P(Y_i \leqslant j)/P(Y_i > j)]$. This model adds effects $\{\beta_i\}$ of the drug dosages on the response to the null model that contains only parameters $\{\alpha_j\}$ pertaining to the logit of each cumulative probability. It treats the effects $\{\beta_i\}$ as identical for each cumulative probability; that is, the effect does not depend on j in the model formula.

This form of model is called *proportional odds*.[44] Independence of dose and response is equivalent to $\beta_1 = \cdots = \beta_I$, each cumulative probability then being identical for all doses. Using a minus sign before the effect of dose in equation (5) implies that the higher the value of β_i relative to other $\{\beta_a\}$, the *lower* the cumulative probability tends to be at dose i, and hence the higher the response tends to be at dose i compared to other doses. The response distributions are stochastically ordered according to $\{\beta_i\}$. The case of a monotone relationship with direction (2) corresponds to $\beta_1 \leqslant \cdots \leqslant \beta_I$.

A monotone relation in which the trend is linear in dose scores $\{d_i\}$ has the simpler model form

$$\text{logit}[P(Y_i \leqslant j)] = \alpha_j - \beta d_i, \quad i = 1, \ldots, I, \; j = 1, \ldots, J-1 \tag{6}$$

with $\beta > 0$ implying (2). The ordinary logistic regression model with a linear dose effect is the special case $J = 2$. For this ordinal model, the odds that the response falls above any given category are multiplied by $\exp(\beta)$ for each unit increase in dose.

The ML fit of any model of this type yields estimated cumulative probabilities at each dose, and hence predicted numbers of observations (fitted values) in the cells of the table. One can test the fit using Pearson or likelihood-ratio chi-squared statistics that compare the observed cell counts to the model's fitted values. The adequacy of these goodness-of-fit tests improves as the cell counts increase in size, the Pearson test being preferred if the cell counts are relatively small.

Model (6) treats the doses as ordinal, whereas model (5) treats them as nominal. To increase power for testing independence when one expects a monotone trend, it is better to use model (6) as the alternative rather than (5). This leads to single degree-of-freedom chi-squared tests for testing independence ($\beta = 0$ in this model).

The likelihood-ratio test has chi-squared statistic given by double the difference in maximized log-likelihoods between the fit of model (6) and the simpler independence model having $\beta = 0$.

Table IV. Example of part of SAS output (using PROC LOGISTIC) for fitting proportional odds model (6) to Table I

Model Fitting Information and Testing Global Null Hypothesis BETA = 0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
− 2 LOG L	2470·961	2461·349	9·612 with 1 DF (p = 0·0019)
Score			9·429 with 1 DF (p = 0·0021)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCP1	1	− 0·7192	0·1588	20·5072	0·0001
INTERCP2	1	− 0·3186	0·1564	4·1486	0·0417
INTERCP3	1	0·6917	0·1579	19·1809	0·0001
INTERCP4	1	2·0570	0·1737	140·2550	0·0001
DOSE	1	− 0·1755	0·0563	9·7087	0·0018

The Wald test is based on the square of the ratio of the ML estimate of β to its standard error. A third test, the 'efficient score' test, is based on the derivative of the log-likelihood function at $\beta = 0$. This score test is equivalent to a generalized CMH correlation test using the dose scores for X and midranks for categories of Y. One can perform all three tests using SAS software for this model, PROC LOGISTIC, as illustrated in Table II. For any of these statistics, one can refer the signed square root (that is, having the same sign as $\hat{\beta}$) to the standard normal distribution to construct a one-sided P-value. PROC LOGISTIC provides the ML fit of the model; PROC CATMOD can also fit the model, but only using WLS. The two approaches give similar results for large samples, but ML is preferred for small samples.

For Table I, the more general model (5) fitted with the constraint $\hat{\beta}_1 = 0$ has estimates $\hat{\beta}_2 = 0{\cdot}118$ (ASE = 0·178), $\hat{\beta}_3 = 0{\cdot}317$ (ASE = 0·175), and $\hat{\beta}_4 = 0{\cdot}521$ (ASE = 0·178). The estimates suggest a monotone increase in response as a function of dose. The likelihood-ratio test that $\beta_2 = \beta_3 = \beta_4 = 0$ has test statistic equal to 9·75 with d.f. = 3 ($P = 0{\cdot}021$). There is evidence of a drug effect, though only the estimate for the high dose level shows substantial evidence of differing from the placebo.

Table IV shows sample SAS output for the simpler model (6), using dose scores (1, 2, 3, 4). One can compare the fit of this model to that of the model (5) with separate dose effects using the difference of −2 log-likelihood values for the two models. Under the hypothesis that the simpler model is adequate, this difference is an approximate chi-squared statistic with d.f. equal to the number of dose levels −2. In this case, the test statistic comparing the models is 2461·35 − 2461·22 = 0·13 with d.f. = 2, indicating that the simpler model is adequate. It has $\hat{\beta} = 0{\cdot}176$ (ASE = 0·056) (SAS actually reports the negative of this value, since it parameterizes the model with + rather than − as the coefficient of the effect). The z test statistics equal $3{\cdot}10 = \sqrt{9{\cdot}612}$ for the likelihood-ratio test, $3{\cdot}12 = \sqrt{9{\cdot}709}$ for the Wald test, and $3{\cdot}07 = \sqrt{9{\cdot}429}$ for the score test, all having a one-sided P-value of 0·001.

These three tests tend to show similar results for large samples. They are valid for smaller samples than one needs for performing goodness-of-fit tests for the model. In fact, even if model (6) does not fit well (as is, in fact, the case for these data), the test statistics provide relatively

powerful tests, compared to tests that ignore the ordering of doses or responses, as long as the linear term in the model represents a major component of the departure from independence. That is, one does not need to test the goodness-of-fit of the model before conducting the association test. In this regard, the remark of Mantel[17] in a similar context is instructive, 'that a linear regression is being tested does not mean that an assumption of linearity is being made. Rather it is that that test of a linear component of regression provides power for detecting any progressive association which may exist.'

Proportional odds models have several appealing properties. If the model holds for a particular set of response categories, it holds with the same parameter effects when the response scale is collapsed in any way. This behaviour is true, approximately, for the sample data. For instance, if we combine the major and minor disability categories in Table I and again fit model (6), we get $\hat{\beta} = 0{\cdot}185$ (ASE = 0·060), compared to $\hat{\beta} = 0{\cdot}176$ (ASE = 0·056) for the complete table. When this model fits well, different studies using different definitions of response categories should reach similar conclusions. In addition, it is unnecessary to assign scores to the response categories.

Once an effect is established, the more complex model (5) is useful for comparing response distributions at different dosage levels. For instance, the difference $\hat{\beta}_i - \hat{\beta}_j$ divided by its standard error is a standard normal test statistic for judging whether the pair of doses i and j is significantly different. One can use Bonferroni methods for simultaneous comparisons; for instance, to simultaneously compare all $(I - 1)$ pairs of adjacent dose levels with an overall type I error probability of no greater than 0·05, one uses nominal size $0{\cdot}05/(I - 1)$ for each pairwise test.

3.2. Other ordinal models

Though the proportional odds model is currently a popular one for modelling ordinal response data, one could alternatively use other ordinal models to detect a monotone dose effect. For instance, McCullagh[44] discussed transforms other than the logit for the cumulative probability, such as the probit and ones (log-log and complementary log-log) for which the cumulative probability approaches 0 at a different rate than it approaches 1. The probit usually provides similar results as the logit, in terms of testing for an effect. McCullagh showed that the probit and logit are most appropriate when an underlying continuous response is roughly bell-shaped, and when a similar form of model holds for that continuum. For instance, if an underlying normal response has approximately a linear relationship with dose, then the logit or the probit of the cumulative probabilities with a linear dose effect tends to fit well. Log-log links, on the other hand, are appropriate when an underlying response is highly skewed. All these options are available with PROC LOGISTIC in SAS.

Other ordinal models utilize single-category probabilities rather than cumulative probabilities. For instance, the *adjacent-categories logit* model with a linear dose effect has form

$$\log[P(Y_i = j)/P(Y_i = j + 1)] = \alpha_j - \beta d_i, \quad i = 1, \ldots, I, \; j = 1, \ldots J - 1. \tag{7}$$

The same dose effect β occurs for logits for each pair of adjacent response categories. Independence is the special case $\beta = 0$, and one can test this with a likelihood-ratio, Wald, or efficient score test. The score test is equivalent to the generalized CMH correlation test using the dose scores for X and equally-spaced scores for the response categories. One can fit models of this form using PROC CATMOD in SAS.[45] With SAS, an ML fit is possible, but it is much simpler to prepare code for the WLS fit; see Table II. Model (7) is also equivalent to an ordinal log-linear

model that uses these scores for the two classifications, called the *linear-by-linear association* model (Agresti,[20] Section 8.1).

The parameters for the effect in models (7) and (6) refer to different types of odds ratios. For instance, $\exp(\beta)$ in model (7) refers to the multiplicative effect of a one-unit increase in dose on the odds of response in the higher instead of the lower of any two adjacent categories. The two models usually fit well in similar situations and provide similar results in the tests.

For instance, with Table I, the estimated effect in model (7) is $\hat{\beta} = 0{\cdot}070$, with standard error 0·023. The z test statistic versions of the Wald and likelihood-ratio statistics equal 3·09 and 3·11, again giving one-tailed P-values of 0·001. The model fits fairly well (Pearson goodness-of-fit statistic = 15·8, d.f. = 11). Both it and the proportional odds model (6) show some lack of fit in the second column for the last row, the response count of 4 in this cell being significantly smaller than the value of nearly 14 that the model predicts.

Finally, the *continuation-ratio* logit model with common effect for each logit has form

$$\log[P(Y_i = j)/P(Y_i \geqslant j + 1)] = \alpha_j - \beta d_i, \quad i = 1, \ldots, I, \; j = 1, \ldots, J - 1. \tag{8}$$

A score test for this model is equivalent to the test of Tarone[43] for survival data discussed in Section 2.5; that is, it is a generalized CMH test based on the sets of probabilities used in these logits. The statistic value and resulting P-value differs from the ones using continuation-ratio logits of form $\log[P(Y_i = j + 1)/P(Y_i \leqslant j)]$.

4. GENERALIZATIONS FOR STRATIFIED DATA

Typically, one studies dose–response relations while controlling for factors that could influence the relationship. For instance, one might display the relationship separately for men and for women, for different age groups, for different centres from which the data are obtained, or for different stages or levels of severity of the medical condition being treated. To illustrate, Table V is a stratified version of Table I that classifies subjects according to the trauma severity at the time of study entry. The study was designed to enroll about the same number of mild versus moderate/severe patients, and the randomization was carried out with severity grade as a stratifying factor. In general, the stratification can be part of the study design or represent post-study control to form more homogeneous subgroups.

For a stratified table, interest focuses not only on the effect of the dose on the response within each stratum, but also on whether there is interaction. Does the dose effect vary according to the stratum?

4.1. Models for the stratified case

The model-building approach can easily accommodate stratified data. We illustrate this with the proportional odds model. For level h of S strata, let Y_{hi} denote a response for a subject at dose level i. The proportional odds model

$$\text{logit}[P(Y_{hi} \leqslant j)] = \alpha_j - \beta_h^S - \beta_i^D, h = 1, \ldots, S, \; i = 1, \ldots, I, \quad j = 1, \ldots, J - 1 \tag{9}$$

has dose effects $\{\beta_i^D\}$ and stratum effects $\{\beta_h^S\}$, but assumes a lack of dose-by-stratum interaction. That is, the effects of the doses on the response are assumed to be the same in each stratum. The special case of this model replacing β_i^D by $\beta^D d_i$ for the dose scores $\{d_i\}$ is relevant for detecting a particular type of monotone trend. One can then construct single-degree-of-freedom

Table V. Data from Table I stratified by trauma severity

Trauma severity	Treatment group	Glasgow Outcome Scale				
		Death	Vegetative state	Major disability	Minor disability	Good recovery
Mild	Placebo	2	4	29	43	26
	Low dose	2	4	25	39	23
	Medium dose	1	3	23	49	24
	High dose	0	1	21	47	26
Moderate/ severe	Placebo	57	21	17	5	6
	Low dose	46	17	19	8	7
	Medium dose	43	11	31	15	7
	High dose	43	3	28	11	15

chi-squared statistics (or, taking square roots, z statistics) testing whether $\beta^D = 0$ using the likelihood-ratio, Wald, or score approaches, in the same way as just discussed for two-way tables.

To illustrate, applying the simpler model with a linear dose effect and dose scores (1, 2, 3, 4) to Table V, we get $\hat{\beta} = 0{\cdot}205$ (ASE = 0·058). The Wald chi-squared statistic equals 12·5, and the likelihood-ratio statistic comparing this model to the simpler one without the dose effect equals the difference in -2-log-likelihood values for the two models, which is also 12·5 ($z = 3{\cdot}53$). The P-value is less than 0·001.

More generally, one could extend model (9) or the simpler one with the linear effect by permitting dose-by-stratum interaction. The model simply then adds cross-product terms of the dose and strata variables (or dummy variables). One can test the hypothesis of no interaction by comparing the -2 log-likelihood values for this model and the corresponding model without interaction. When the degree of interaction seems substantively important, one can estimate and test the effect separately in each stratum using the dose effect estimates pertaining to that stratum, or one could simply fit the original model (for example, 6) separately to each stratum to obtain the separate effects. (This approach is not equivalent, because it estimates intercept parameters separately with each fit.) On the other hand, when the dose effects do not vary much among the strata, the overall test based on a lack of interaction tends to be much more powerful, and the overall estimate tends to be more efficient, since they summarize information across the strata.

In fact, there is some evidence of interaction in Table V. For the models with linear dose effect, the likelihood-ratio statistic comparing the model with separate slopes to the model with a single slope equals 3·85 with d.f. = 1. The model with separate slopes has estimates $\hat{\beta} = 0{\cdot}099$ (ASE = 0·082) for the mild trauma group and $\hat{\beta} = 0{\cdot}327$ (ASE = 0·082) for the moderate/severe trauma group. Hence, there is a strong evidence of a dose effect only for the latter group. There are other approaches one could use both to check for interaction and to describe the separate effects, but we do not discuss them here because of space limitations.

4.2. Non-model-based approaches for the stratified case

The CMH approach generalizes naturally to combining information from several strata; in fact, the original statistic presented by Mantel and Haenszel[46] was designed specifically for the stratified case with two groups and a binary response. For the case of several doses and an ordinal

response, the correlation statistic (Mantel[17]) provides a large-sample chi-squared statistic with d.f. = 1 for detecting a linear trend in the effect. One can, as usual, treat the signed square root as a standard normal statistic. The CMH approach, like model (9), works well when the dose effects are similar in each stratum. It is available with the CMH1 option in PROC FREQ in SAS. For Table V, this approach used with equally-spaced scores for doses and response outcomes yields a chi-squared statistic of 16·2 and normal statistic of 4·0, for which the P-value is less than 0·001.

Similarly, one could consider stratified versions of tests discussed in Section 2 that are special cases of a generalized CMH test, such as Tarone's test.[42] In principle, this type of construction could also be used with other sorts of statistics, such as the Jonckheere-Terpstra statistic.

5. SMALL-SAMPLE AND SPARSE-DATA INFERENCE

The test statistics presented in this article are large-sample statistics. For chi-squared statistics, the convergence to chi-squared distributions tends to be faster for statistics having smaller values of d.f., such as the single-degree-of-freedom statistics.

For any particular statistic referring to the two-way contingency table of dose by ordinal response, one can construct a small-sample 'exact' test using the generalized hypergeometric distribution that results from conditioning on the row and column totals. This approach generalizes Fisher's exact test for 2-by-2 tables, with the conditioning argument yielding a distribution not depending on unknown nuisance parameters. Exact tests are available in StatXact[22] for several statistics, including the Jonchkeere–Terpstra statistic and correlation-type statistics with fixed or rank scores.[47] (The correlation-type statistics use the 'linear-by-linear' option in StatXact.) Currently these tests are restricted to the single-stratum case.

For stratified data, only the case of two dose groups is currently addressed by standard software, using the CMH correlation type approach for a set of fixed or midrank response scores (StatXact). In principle, though, the methodology of small-sample exact tests extends directly to the more general case of several dose groups.[48] Specialized FORTRAN programs exist for these analyses.

6. SAMPLE SIZE AND POWER

Whitehead[49] discussed sample size formulae for an ordered categorical response with the proportional odds model, though only for the case of two groups (for example, two doses). Suppose we want power $1 - \beta$ in an α-level test for detecting an effect of size β_0 in that model. The sample is to be allocated to the two groups in the ratio A to 1, and $\bar{p}_j$ denotes the anticipated marginal proportion in response category j. Whitehead[49] stated that the required sample size for a two-sided test is then approximately

$$N = 3(A + 1)^2(z_{\alpha/2} + z_\beta)^2/[A\beta_0^2(1 - \sum \bar{p}_j^3)]$$

where z_a is the $100(1 - \alpha)$ percentile of the standard normal distribution.

This requires anticipating the marginal proportions as well as the size of the effect. Setting $\bar{p}_j = 1/J$ provides a lower bound for N. Whitehead[49] showed that the sample size does not depart much from this bound unless a single dominant response category occurs. Hilton and Mehta[50] provided a somewhat different approach to sample size determination, based on evaluating the exact conditional distribution with a network algorithm, or simulating that distribution.

With equal marginal probabilities, Whitehead's[49] formula is useful for showing the effect of the choice of number of response categories. The ratio of the sample size $N(J)$ needed for J categories relative to the sample size $N(2)$ needed for two categories is

$$N(J)/N(2) = 0{\cdot}75/[1 - 1/J^2].$$

Relative to a continuous response ($J = \infty$), using J categories provides efficiency $(1 - 1/J^2)$. The loss of information from collapsing to a binary response is substantial, but there is little gain from using more than about five categories. For fixed J, equal allocation ($A = 1$) produces the smallest sample size.

The case of $I > 2$ groups does not seem to have been considered. However, various rather *ad hoc* ways exist of approaching the problem. For instance, many tests discussed in this article are based on asymptotically normal statistics, such as a measure of association (for example, correlation, gamma) or an estimate of a model parameter ($\hat{\beta}$ for the proportional odds model). Let $\hat{\theta}$ denote a generic asymptotically normal estimator of a parameter θ, with variance of the form V/N. Then, for a fixed non-null value θ_0 of θ, standard arguments show that the required sample size for a one-sided test is

$$N = (z_\alpha + z_\beta)^2 V/\theta_0^2.$$

To use this formula, the steps are to: (i) choose an anticipated set of non-null cell probabilities; (ii) find the value of θ_0 corresponding to those probabilities; (iii) find V for those probabilities, and (iv) substitute V into this formula using the required size and power. In some cases V has closed form, based on the delta method, and in some cases it requires iterative methods. Even when it has closed form, though, the formula is typically messy computationally. A simple approach to determining V (and θ_0) enters the anticipated probabilities as data into standard software, in which case V equals the square of the reported asymptotic standard error.

For illustrative purposes, suppose we had anticipated probabilities proportional to the counts in Table I. For the proportional odds model, we observed $\hat{\beta} = 0{\cdot}1755$ and a standard error of 0·0563 for these data having a sample size of 802. Setting $V/802 = (0{\cdot}0563)^2$ yields $V = 2{\cdot}542$. To have power 0·90 in an $\alpha = 0{\cdot}05$ level one-sided test of $\beta = 0$ when the true relationship has $\beta_0 = 0{\cdot}1755$ requires a sample size of about $N = (1{\cdot}645 + 1{\cdot}282)^2(2{\cdot}542)/(0{\cdot}1755)^2 = 707$.

For stratified data, Whitehead[49] noted that logistic regression and an ordinal extension such as the proportional odds model may require a somewhat increased sample size to preserve the desired power. However, the variation among strata in the category probabilities has to be quite extreme before sample size is greatly affected.

SUMMARY AND RECOMMENDATIONS

We have presented a variety of tests for detecting a monotone relation between dose and an ordinal response. Of the non-model-based methods, the tests based on the correlation seem the most flexible. These connect closely with methods used for continuous reponses, which can be regarded as a limiting case as the number of response categories and doses increases indefinitely. Though directed toward a narrow alternative, namely linearity for the choice of scores, this provides the advantage of good power if a strong linear component exists for the true association.

The order-restricted approach has the advantage of specifying the alternative in a broader and more realistic manner. Disadvantages include a rather awkward limiting distribution, a lack of a full theoretical and methodological development for a categorical response when the number of

doses exceeds two or the data are stratified, and potential power loss compared to a linear trend statistic when the true relation has a strong linear component. Some preliminary power studies by one of the authors for a separate project suggest that the order-restricted approach has better power than the linear trend test if the response is essentially identical for all positive dose groups but those groups have better response than the control group. On the other hand, for other patterns of monotone increase that do not depart so drastically from linearity, the linear trend statistic is more powerful.

Section 2 addressed the dose–response relationship within the significance-testing framework. Our overall preference, however, is for a model-based approach, since it provides a fuller description of the dose–response relationship. For instance, estimated odds ratios describe the strength of the effect, and fitted values provide estimates of response probabilities that are smoother and tend to have smaller mean squared errors than the sample proportions. Goodness-of-fit tests check the model adequacy, and residuals can indicate potential departures from the trend predicted by the model. Moreover, the fit of a model such as (5) enables us to consider the more important follow-up questions, such as determining which doses have significantly different responses and which dose is optimal.

Some statisticians avoid the model-building approach for fear of increasing the number of assumptions, with the resulting test being less robust. However, many of the standard tests have connections with models, being efficient score tests. For large samples, one obtains similar results from a likelihood-ratio test for a model parameter as one does from a score test. For Table I, for instance, model-based and non-model-based tests gave similar results.

Focusing on models makes one recognize the structure under which a particular test is natural. Moreover, a model-based test can provide a powerful approach even if the model does not fit well. For instance, for testing conditional independence in stratified tables, generalizations of the Cochran–Mantel–Haenszel test are popular. These tests are score tests for models that assume homogeneity of odds ratios across strata. However, one does not need to assume such homogeneity to use the tests, and they perform well whenever the true degree of heterogeneity is not severe.

Finally, emphasizing models has the advantage of decreasing reliance on significance tests as the primary mode of analysis. Though this paper has surveyed a variety of such tests for detecting monotone dose-response relationships, ultimately estimation of parameters yields more informative conclusions.

ACKNOWLEDGEMENTS

The work of Agresti was partially supported by an NIH grant.

REFERENCES

1. Hothorn, L. 'Robustness study on Williams- and Shirley- procedure, with application in toxicology', *Biometrical Journal*, **31**, 891–903 (1989).
2. Piegorsch, W. W. 'Nonparametric methods to assess non- monotone dose response: Applications to genetic toxicity', in Sen, P. K. and Salama, I. A. (eds.), *Order Statistics and Nonparametrics: Theory and Applications*, Elsevier Science Publishers, B. V., 1992.
3. Dalal, S. N. and Lawson, J. S. 'Methods for the dose response analysis in animal growth promotion studies', *ASA Proceedings of the Biopharmaceutical Section*, 171–176 (1989).
4. Maclure, M. and Greenland, S. 'Tests for trend and dose response: Misinterpretations and alternatives', *American Journal of Epidemiology*, **135**, 96–104 (1992).
5. Breslow, N. E. and Day, N. E. 'Statistical methods in cancer research', in *The Design and Analysis of Cohort Studies*, Vol. 2, International Agency for Research on Cancer, 1987, p. 97.

6. Mack, G. A. and Wolfe, D. A. 'K-sample rank tests for umbrella alternatives', *Journal of the American Statistical Association*, **76**, 175–181 (1981).
7. Hettmansperger, T. P. and Norton, R. M. 'Tests for patterned alternatives in k-sample problems', *Journal of the American Statisical Association*, **82**, 292–299 (1987).
8. Chen, Y. I. and Wolfe, D. A. 'Modifications of the Mack–Wolfe umbrella tests for a generalized Behrens–Fisher problem', *Canadian Journal of Statistics*, **18**, 245–253 (1990).
9. Chen, Y. I. and Wolfe, D. A. 'A study of distribution-free tests for umbrella alternatives', *Biometrical Journal*, **32**, 47–57 (1990).
10. Chen, Y. I. 'On the comparison of umbrella pattern treatment means with a control mean', *Biometrical Journal*, **35**, 689–700 (1993).
11. Ruberg, S. J. 'Dose-response studies. I. Some design considerations', *Journal of Biopharmaceutical Statistics*, **5**, 1–14 (1995).
12. Ruberg, S. J. 'Dose-response studies. II. Analysis and Interpretation', *Journal of Biopharmaceutical Statistics*, **5**, 15–42 (1995).
13. Kodell, R. L. and Chen, J. J. 'Characterization of dose-response relationships inferred by statistically significant trend tests', *Biometrics*, **47**, 139–146 (1991).
14. Zeger, S. L. and Liang, K. Y. 'Dose-response estimands. Comment on "Compliance as an explanatory variable in clinical trials"', *Journal of the American Statistical Association*, **86**, 18–19 (1991).
15. Yates, F. 'The analysis of contingency tables with grouping based on quantitative characters', *Biometrika*, **35**, 176–181 (1948).
16. Armitage, P. 'Tests for linear trends in proportions', *Biometrics*, **11**, 375–386 (1955).
17. Mantel, N. 'Chi-square tests with one degree of freedom: extensions of the Mantel–Haenszel procedure', *Journal of the American Statistical Association*, **58**, 690–700 (1963).
18. Graubard, B. I. and Korn, E. L. 'Choice of column scores for testing independence in ordered 2XK contingency tables', *Biometrics*, **43**, 471–476 (1987).
19. Landis, J. R., Heyman, E. R. and Koch, G. G. 'Average partial association in three-way contingency tables: A review and discussion of alternative tests', *International Statistical Review*, **46**, 237–254 (1978).
20. Agresti, A. *Categorical Data Analysis*, Wiley, 1990.
21. Jonckheere, A. R. 'A distribution-free K-sample test against ordered alternatives', *Biometrika*, **41**, 133–145 (1954).
22. StatXact. *StatXact3 for Windows: Statistical Software for Exact Nonparametric Inference, User Manual*, Cytel Software, 1995.
23. Heeren, T. and D'Agostino, R. 'Robustness of the two independent samples *t*-test when applied to ordinal scaled data', *Statistics in Medicine*, **6**, 79–90 (1987).
24. Grizzle, J. E., Starmer, C. F. and Koch, G. G. 'Analysis of categorical data by linear models', *Biometrics*, **25**, 489–504 (1969).
25. Bartholomew, D. J. 'Ordered tests in the analysis of variance', *Biometrika*, **48**, 325–332 (1961).
26. Bartholomew, D. J. 'A test of homogeneity of means under restricted alternatives', *Journal of the Royal Statistical Society, Series B*, 239–281 (1961).
27. Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. *Statistical Inference Under Order Restrictions*, Wiley, 1972.
28. Robertson, T., Wright, F. T. and Dykstra, R. L. *Order- Restricted Statistical Inference*, Wiley, 1988.
29. Brunden, M. N. 'A review of isotonic regression theory – means from normal distribution', TR 9164-95-001, The Upjohn Co., Kalamazoo, MI, 1995.
30. Williams, D. A. 'A test for differences between treatment means when several dose levels are compared with a zero dose control', *Biometrics*, **27**, 103–117 (1971).
31. Williams, D. A. 'The comparison of several dose levels with a zero dose control', *Biometrics*, **28**, 519–531 (1972).
32. Capizzi, T., Survill, T. T., Heyse, J. F. and Malani, H. 'An empirical and simulated comparison of some tests for detecting progressiveness of response with increasing doses of a compound', *Biometrical Journal*, **34**, 275–289 (1992).
33. Tukey, J. W., Ciminera, J. L. and Heyse, J. F. 'Testing the statistical certainty of a response to increasing doses of a drug', *Biometrica*, **41**, 295–301 (1985).
34. Bartholomew, D. J. 'A test of homogenity for ordered alternatives', *Biometrika*, **46**, 36–48 (1959).

35. Shirley, E. 'A non-parametric equivalent of Williams' test for contrasting increasing dose levels of a treatment', *Biometrics*, **33**, 386–389 (1977).
36. Cohen, A. and Sackrowitz, H. B. 'Improved tests for comparing treatments against a control and other one-sided problems', *Journal of the American Statistical Association*, **87**, 1137–1144 (1992).
37. Hayter, A. J. 'A one-sided Studentized range test for testing against a simple ordered alternative', *Journal of the American Statistical Association*, **85**, 778–785 (1990).
38. Silvapulle, M. J. and Silvapulle, P. 'A score test against a one-sided alternative', *Journal of the American Statistical Association*, **90**, 342–349 (1995).
39. Grove, D. M. 'A test of independence against a class of ordered alternatives in a $2 \times c$ contingency table', *Journal of the American Statistical Association*, **75**, 454–459 (1980).
40. Robertson, T. and Wright, F. T. 'Likelihood-ratio tests for and against a stochastic ordering between multinomial populations', *Annals of Statistics*, **9**, 1248–1257 (1981).
41. Patefield, W. M. 'Exact tests for trends in ordered contingency tables', *Journal of the Royal Statistical Society, Series C*, **31**, 32–43 (1982).
42. Grove, D. M. 'Positive association in a two-way contingency table: Likelihood ratio tests', *Communications in Statistics, A: Theory and Methods*, **13**, 931–945 (1984).
43. Tarone, R. E. 'Tests for trend in life table analysis', *Biometrika*, **62**, 679–682 (1975).
44. McCullagh, P. 'Regression models for ordinal data (with discussion)', *Journal of the Royal Statistical Society, Series B*, **42**, 109–142 (1980).
45. Stokes, M. E., Davis, C. S. and Koch, G. G. *Categorical Data Analysis Using the SAS System*, SAS Institute Inc, 1995.
46. Mantel, N. and Haenszel, W. 'Statistical aspects of the analysis of data from retrospective studies of disease', *Journal of the National Cancer Institute*, **22**, 719–748 (1959).
47. Agresti, A., Mehta, C. R. and Patel, N. R. 'Exact inference for contingency tables with ordered categories', *Journal of the American Statistical Association*, **85**, 453–458 (1990).
48. Kim, D. and Agresti, A. 'Nearly exact tests of conditional independence and marginal homogeneity for sparse contingency tables', *Computational Statistics and Data Analysis*, **24**, 89–104 (1997).
49. Whitehead, J. 'Sample size calculations for ordered categorical data', *Statistics in Medicine*, **12**, 2257–2271 (1993).
50. Hilton, J. F. and Mehta, C. R. 'Power and sample size calculations for exact conditional tests with ordered categorical data', *Biometrics*, **49**, 609–616 (1993).

Index

Index complied by Geoffrey C. Jones